UNIVERSITY OF CAMBRIDGE

Evaluating Natural Language Generation Tasks for Grammaticality, Faithfulness and Diversity

Huiyuan Xie

Downing College

This dissertation is submitted on 30 June 2023 for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60 000 words.

Huiyuan Xie
June 2023

Abstract

Natural language generation (NLG) plays a vital role in many applications. Evaluating the quality of generated text is crucial for ensuring the effectiveness and user satisfaction of NLG systems. With the popularisation of deep learning in recent years, many models have been reported to achieve super-human performance on popular benchmarks. However, it has been observed that existing holistic benchmarks and evaluation metrics frequently fail to accurately assess specific evaluation factors that are of interest to the field.

This thesis explores a diagnostic evaluation framework for assessing the grammaticality, faithfulness, and diversity (GFD) of generated text in NLG tasks. These three metrics are considered as essential linguistic qualities, which need to be present in the outputs of NLG models. Grammaticality is examined by analysing the parsability of a sentence with a well-defined formal grammar. Faithfulness is divided into two facets: grounding faithfulness and task faithfulness. These two facets investigate how well the model outputs align with both the information provided in the input, and the inherent requirements of the task. Diversity is further divided into word-level and parse-level diversity measures. In the proposed GFD framework, the evaluation of the three metrics does not require task-specific references to be constructed. By clearly defining and evaluating these generation qualities, this framework aims to provide insights into the strengths and limitations of NLG models.

To demonstrate the versatility of the GFD evaluation framework, three different generation tasks are explored: synthetic image captioning, football highlight generation from match statistics, and topic-shift dialogue generation. These tasks are deliberately chosen to cover a diverse range of generation scenarios. Each task provides unique grounding information and constraints that influence the generation process, which in turn create diverse challenges for the evaluation of NLG models. Experiments on these tasks reveal the challenges in fine-grained NLG evaluation when the availability of ground truth representations diminishes or when there is a delicate balance between input groundings and task constraints. This thesis empirically demonstrates how the GFD evaluation framework, in combination with diagnostic datasets, can provide insights into model strengths and limitations to supplement standard evaluations.

Acknowledgements

I would like to express my sincere gratitude to all those who have supported and guided me throughout my journey of completing this PhD thesis.

First and foremost, I would like to express my deepest gratitude to my supervisor, Ann Copestake. Her exceptional guidance, expertise, and encouragement have been instrumental in shaping the outcome of this research. I am truly grateful for her insightful feedback and constant support, which I will continue to benefit from.

I extend my sincere thanks to all my collaborators who have contributed to this research. I would like to thank Alexander Kuhnle for his assistance during my first year, patiently answering all my trivial questions about ShapeWorld. I am grateful to Chenyan Xiong's valuable suggestions on my TIAGE work. Additionally, I am thankful to the annotators who contributed to the football templates and topic-shift labels. It has been an honour to collaborate with such exceptional individuals.

I would like to express my sincere gratitude to my viva examiners, Paula Buttery and Dan Flickinger, for their probing questions and insightful comments, which have significantly enriched the quality of this research. I would like to thank Simone Teufel, Andreas Vlachos and Weiwei Sun for the insightful feedback they provided during my first-year and third-year exams. I am grateful to Guy Emerson for his valuable comments on the grammaticality work, and to Dan Flickinger for the assistance with ERG relaxation.

I would like to acknowledge my colleagues and friends, including Tianqi Huang, Sancia Xie, Molly Xia, Yimai Fang, Meng Zhang, James Thorne, Lily Li, Sherry Yin, Linshu Feng, Yuki Wang and many others for their support, camaraderie and friendship.

I would like to express my heartfelt appreciation to Jack, who has been my rock and my greatest cheerleader throughout this challenging yet rewarding endeavour. His unwavering support and the numerous "you can do it" cards have provided me with the encouragement and motivation I needed during difficult times. I am immensely grateful for his belief in me and his constant presence by my side. Furthermore, I would like to extend my heartfelt thanks to the whole Bradley family for their warmth and kindness. Throughout this journey, they have helped me in so many ways. I would also like to express a particular thank you to little Lyla, whose infectious smile brought much-needed light and positivity to even the most trying days of writing.

Finally, I want to express my deepest gratitude to my family. Their unconditional love, encouragement and belief in my abilities have been the driving force behind my pursuit of knowledge. I am forever grateful for their unwavering support.

Contents

Chapter 1

Introduction

Over recent years, the field of natural language generation (NLG) has experienced remarkable progress, thanks to rapid advancements in machine learning and deep learning techniques, coupled with the availability of large datasets. These advancements have facilitated the development of sophisticated models that can generate human-like text. The increasing complexity and capabilities of these systems have simultaneously amplified the challenge of effectively evaluating their performance. Considerable efforts have been devoted to establishing efficient and straightforward evaluation practices (Hardcastle and Scott, 2008; Post, 2011; Hastie and Belz, 2014; Lau et al., 2015; Liu et al., 2016; Zhu and Bhat, 2020; Sun et al., 2022). Typical evaluation methods in NLG usually consist of an *automatic evaluation* employing generic metrics on a pre-constructed test set (Papineni et al., 2002; Lin, 2004; Vedantam et al., 2015; Doddington, 2002; Anderson et al., 2016; Zhang et al., 2019b), complemented by a *human evaluation*, which is often carried out on a smaller, randomly selected subset of the test data (van der Lee et al., 2019; Hämäläinen and Alnajjar, 2021; van der Lee et al., 2021).

Automatic evaluation metrics used for NLG tasks typically provide a holistic score that represents overall model performance. These holistic metrics, referred to as "black-box" evaluation in Dale and Mellish (1998), lack clarity regarding the specific qualities they actually measure. For instance, widely used n-gram overlapping metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), focus solely on lexical-level similarity between model predictions and reference texts, displaying weak correlation with human judgement. Furthermore, most existing metrics rely on a limited number of references to evaluate model-generated sentences, failing to address the inherent challenge of NLG evaluation, where multiple valid predictions may lack extensive lexical or semantic similarity with the references. Existing metrics also fall short of indicating whether model-generated predictions are consistent with the context on which the generation is conditioned, or the strengths and disadvantages of a particular model. As such, while these metrics greatly facilitate automatic evaluation and expedite AI model development, they often fail to shed

light on models' specific strengths and limitations.

Whilst typically considered as the ultimate measure of system performance, human evaluation is susceptible to design deficiencies. Recent studies (van der Lee et al., 2019; Howcroft et al., 2020) have exposed shortcomings in human evaluations reported in NLG research, including imprecise specifications, lack of clarity, and inconsistencies (discussed in detail in section 2.3.1). These deficiencies compromise the effectiveness of human evaluation in calibrating model performance, which is one of the primary objectives of evaluation. Confusion regarding terminology usage in existing literature compounds this issue (Howcroft et al., 2020), where different evaluation frameworks employ the same terminology to describe different aspects of text quality, or use varied terms to refer to the same aspect. The unfortunate end result of design failure and confused terminology is that comparing human evaluation results across different studies becomes extremely challenging.

The recent surge in popularity of large language models (LLMs), such as ChatGPT (OpenAI, 2022), has brought to light the impressive abilities of language models in a multitude of real-world applications. However, in-depth examinations (Li, 2023; Bang et al., 2023) have uncovered that ChatGPT often produces answers that initially appear to be highly plausible; yet on closer examination, these responses are found to lack accuracy and validity. This is unsurprising given that neural language models are known to be prone to hallucination (Rohrbach et al., 2018; Lee et al., 2019; Ji et al., 2022), where models generate seemingly correct sentences that are actually inconsistent with background information or common knowledge.

This highlights the pressing necessity to rethink what evaluation means in the context of NLG, particularly in the era of large models and neural technologies. The crucial question to ask is, is pursuing an all-round, holistic evaluation the ideal direction for NLG evaluation attempts? This thesis argues that while the concept of having a universal and readily deployable evaluation approach may appear appealing, it is crucial to exercise caution regarding the potential compromise of quality in pursuit of efficiency. At the heart of this matter is the recognition that the conceptualisation and implementation of evaluation ultimately requires a case-by-case consideration. It is essential that evaluation metrics align closely with the intrinsic nature of the task and maintain relevance to the focal points of what one wants to measure. Metrics that deviate from these principles, even if they provide observational insights through experimental efforts, fail to serve as valid means of assessing and validating models. These principles apply not only to automatic evaluation practices, but also to the standards of human evaluation.

As a way of making progress on clarifying the issues, I advocate a "glass-box" evaluation approach, and suggest a **decomposition** of the problem of NLG evaluation and a conceptualisation of individual measures within this process.

This approach aligns with the spirit of evaluation efforts in the early days of NLP research. Galliers and Spärck Jones (1993) extensively questioned the evaluation of NLP systems. Their work primarily centred around evaluation of generic NLP systems, advocating a grid organisation which decomposes evaluation into considering both environmental variables and system parameters. Separately, Mellish and Dale (1998) touched upon the question in the particular context of NLG, where the generation process is conventionally pipelined by component tasks including content determination, text structuring and surface realisation. They advocated for a "glass-box" evaluation approach, which considers the problem of NLG evaluation by looking at the evaluation of individual component sub-tasks. This call for decomposing evaluation into component sub-tasks has been echoed in multiple other studies (Dale and Mellish, 1998; Dale et al., 1998).

However, the architecture of generation models has transitioned from the earlier pipeline-based design to a neural paradigm that functions in an *end-to-end* manner. As a result, the particular manner of decomposing NLG evaluation into examination of individual sub-tasks (Mellish and Dale, 1998) has fallen out of favour in the field. Nonetheless, it is my view that the **decomposition** approach and the mindset it entails for evaluation remains valuable. NLP evaluation is inherently a customised matter, as emphasised in the classic publication by Spärck Jones (1994, pp. 104-105):

> It is evident that serious evaluation requires a well-understood *decomposition* of the whole, in terms of the evaluation aims or remit, and design, and the precise definition of the evaluation subject. This applies even in the case where all that is wanted is some snapshot of the performance of an operational system. [...] Therefore, the methodology required for an NLP evaluation is an *unpacking* one, designed to address the very many distinctions involved and make properly-related choices on each.

1.1 Core of the thesis

This thesis is about NLG evaluation. This thesis explores a **decomposition** approach to evaluation, dividing the problem into fundamental language quality measures, and persists in the utilisation of clearly characterised and operationalised terminology to assess the qualities concerned.

This thesis aims to achieve several **key objectives** in the field of NLG evaluation. Firstly, it seeks to provide an overview of evaluation aspects that have been discussed in the existing literature, and clarify the commonly used terminologies in NLG evaluation. Secondly, this thesis aims to investigate the specific aspects to which NLG evaluation should be decomposed. By analysing the various dimensions and components of NLG systems, the study aims to identify and determine the crucial factors for evaluating the

performance and effectiveness of NLG models. Furthermore, it proposes an evaluation protocol that addresses the identified factors. The objective is to develop a structured framework that encompasses the essential qualities to be considered during the evaluation process. Each factor will be characterised and defined to ensure clarity and consistency in the evaluation procedure. Moreover, the research applies the proposed evaluation protocol to distinct NLG generation tasks. By conducting evaluations on different scenarios, the study explores the extent to which the evaluation factors can be generalised, and their applicability across various NLG domains.

To accomplish the aforementioned objectives, several **research questions** will guide the investigation.

1. Which evaluation practices are considered useful in the context of NLG?
2. What qualities should be deemed essential and included in the proposed evaluation protocol? How should these qualities be conceptualised and operationalised?
3. Is it feasible to conduct a comprehensive examination of these identified qualities?
4. How generalisable is the evaluation protocol across different NLG scenarios?
5. What insights can the evaluation with the proposed protocol reveal about NLG models?

This thesis aims to address the above questions, by exploring a principled, well-established evaluation framework for grammaticality, faithfulness and diversity (GFD). By applying the GFD protocol to various NLG generation tasks, the study will assess the protocol's adaptability and its ability to provide meaningful insights across different domains.

The GFD evaluation framework offers several benefits that contribute to a comprehensive and adaptable approach for evaluating language generation. Firstly, the framework examines distinct aspects of evaluation, each focusing on specific syntactic or content-related criterion associated with language generation. By considering multiple dimensions, namely grammaticality, diversity, and faithfulness, the evaluation framework provides a sound assessment of generated outputs.

One key advantage of the GFD framework lies within its clear definition and careful conceptualisation. It offers a well-defined structure that facilitates consistent and rigorous evaluation. However, the framework also allows for necessary tailoring to accommodate individual tasks. This flexibility ensures that the evaluation can be customised to fit the unique requirements and characteristics of different language generation tasks.

Furthermore, the core design of the GFD framework maintains consistency across tasks, alongside minor adjustments made to address task-specific considerations. This

consistency, particularly evident in the evaluation of grammaticality and diversity, enables meaningful cross-task comparisons.

At the same time, the GFD framework also recognises the need for finer adjustments, primarily in relation to the *faithfulness* measure. This ensures the alignment of the evaluation design with the specific requirements of individual tasks. It emphasises the importance of assessing language generation on a case-by-case basis, considering the unique nuances and characteristics of each distinct task.

Overall, the GFD evaluation framework strikes a balance between a well-defined structure and the necessary adaptability for diverse language generation tasks. It offers a robust and consistent evaluation methodology while accommodating task-specific considerations.

It is worth clarifying the following aspects within the confines of this thesis: 1. The linguistic discussions and empirical analyses presented in this study pertain specifically to the English language, which may not be perfectly applicable or transferable to other distinct languages; and 2. The discussions in this thesis concern text-based generation tasks, and do not involve speech formats or other alternative forms of language.

1.2 Thesis contributions

1.2.1 Key contributions

In this thesis, I examine the research question of diagnostic evaluation of natural generation tasks for grammaticality, faithfulness and diversity. The key contributions of this thesis are listed as follows:

- I examine the importance of fine-grained, separate, disentangled evaluation metrics for natural language generation tasks, which look beyond surface n-gram similarity, and inspect model outputs from well-defined perspectives with finer granularity. On top of this discussion, I explore how this evaluation can be achieved computationally.

- I propose a novel, reference-less and fine-grained evaluation framework for evaluating natural language generation models for grammaticality, faithfulness and diversity. The three generation qualities are chosen to be as orthogonal as possible, and are carefully conceptualised with motivations obtained from deep linguistic analysis and task grounding and completion requirements. I also provide detailed accounts of existing works that discussed the three qualities, and provide detailed explanations regarding what is being evaluated for each of the qualities in this thesis.

- I construct three specific datasets, ShapeWorldICE, FHiG and TIAGE, to investigate the GFD evaluation protocol under different generation requirements and constraints.

- I apply the GFD evaluation framework to three generation tasks: synthetic image captioning, data-to-text football highlight generation, and topic-shift dialogue generation. Experiments with the three tasks serve as a *proof-of-concept* for the GFD evaluation paradigm, showcasing its versatility across a variety of task scenarios.

1.2.2 Minor contributions in separate chapters

In addition to the key contributions listed above, there are also minor contributions that spread over the main text of this thesis, and can hopefully serve as useful research tools for the wider research community.

- I provide an extensive survey of assessment approaches for human-written text and model-generated text, and a comprehensive overview of existing dialogue systems.

- I have open-sourced the TIAGE dataset, and scripts that were used to pre-process the data, run baseline experiments and carry out GFD evaluation. I will also release the ShapeWorldICE dataset and code after code refactoring. I also intend to open-source the code that I used to collect the FHiG data, and the manually prepared football highlight templates. This is in the hope that the release of this data and code for public access can facilitate the research in this field.

1.2.3 Publications

First authored, thesis-related publications:

- Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. TIAGE: A benchmark for topic-shift aware dialog modeling. In *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1684–1690, 2021

- Huiyuan Xie, Tom Sherborne, Alexander Kuhnle, and Ann Copestake. Going beneath the surface: Evaluating image captioning for grammaticality, truthfulness and diversity. In *Proceedings of the MetaEval Workshop at the AAAI conference on artificial intelligence (AAAI)*, 2020

First authored publications (not directly related to thesis):

- Huiyuan Xie and Ignacio Iacobacci. Audio visual scene-aware dialog system using dynamic memory networks. In *Proceedings of the DSTC8 Workshop at the AAAI conference on artificial intelligence (AAAI)*, 2020

Co-authored publications (not directly related to thesis):

- Aishan Liu, Huiyuan Xie, Xianglong Liu, Zixin Yin, and Shunchang Liu. Revisiting audio visual scene-aware dialog. *Neurocomputing*, 496:227–237, 2022

- Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive GAN for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 33, pages 1028–1035, 2019

1.3 Thesis outline

In chapter 2, I discuss text quality in the sense of human-written text and model-generated text. I then move on to give an overview of NLG tasks, datasets, models and evaluation methods proposed in recent years. This is to lay a broad foundation for the three evaluation metrics and the three generation tasks that this thesis is most concerned with.

Chapter 3 introduces the core contribution of this thesis - a principled evaluation framework for NLG tasks with an explicit focus on grammaticality, faithfulness and diversity (GFD). I explain the rationale for selecting GFD as the evaluation metrics, followed by a comprehensive conceptualisation of the three metrics. This chapter lays the groundwork for the evaluation attempts set out in subsequent chapters, outlines the insights that can be derived from the GFD framework, and justifies the feasibility of a controlled, diagnostic and fine-grained evaluation practice for NLG scenarios.

Chapter 4 sets the scene for the rest of the thesis. In this chapter, I describe in detail the three NLG tasks that are inspected in this thesis. I explain the rationale behind the selection of the three tasks: synthetic image captioning, football highlight generation, and topic-shift dialogue generation. I elaborate on my decision to select image captioning as the starting point for GFD evaluation, as well as why I specifically opted for a synthetic setup to investigate the evaluation of this task. Following this, I discuss the reasoning behind selecting football highlight generation as a particular case of data-to-text generation, as well as the decision to explore topic-shift dialogue generation for dialogue systems. I also justify how the other two tasks, football highlight generation and topic-shift dialogue generation, present natural next-level challenges for exploration following the image captioning task.

Chapter 5 presents the first NLG task that I investigate in this thesis: synthetic image captioning. I begin by providing an introduction to this task, with an overview of existing work related to the task. I then introduce the ShapeWorldICE benchmark created for this work in section 5.3, explore the feasibility of utilising the GFD framework to evaluate the synthetic image captioning task, and present evaluation results, alongside detailed discussion.

In chapter 6, I look into another generation task - **F**ootball **H**ighl**i**ght **G**eneration (FHiG). I consider this task an appropriate next step for investigation following the

ShapeWorldICE evaluation work, as the FHiG task bridges the requirements for a controlled generation setup and a real-world scenario, due to the semi-structured nature of its input. In this chapter, I start off by providing a formal definition of the task in section 6.1. I then introduce the specifically curated dataset for this task in section 6.2. Similar to the ShapeWorldICE work, a GFD evaluation framework is proposed for the football highlight generation task. Baseline model implementation and experiment results are presented, followed by a discussion of the insights obtained from the evaluation practice.

Chapter 7 expands on the use of the GFD protocol for NLG evaluation to another scenario: open-domain chit-chat dialogue generation with a specific focus on topic shifting. I introduce the task of topic-shift dialogue generation in section 7.1, followed by a thorough discussion of the TIAGE dataset specifically curated for this task. The baseline models utilised in this task are introduced in section 7.4. Furthermore, I test my GFD framework on this particular task, investigating the feasibility and versatility of the GFD evaluation in an open-domain chit-chat environment.

Chapter 8 concludes this thesis by reflecting on the work presented in this thesis, highlighting the main contributions, and outlining potential future directions towards better evaluation practices for NLG.

Chapter 2

Background and motivation

Linguistic research has long been exploring properties of human texts in a wide range of use cases. The advent of text generation technologies generated a new body of natural language texts from computational systems. In the literature there have been evaluation attempts looking at the quality of both human-written texts and model-generated texts. The two sources of texts (*human-written* vs *model-generated*) and the two evaluation methods (*human assessment* vs *automated assessment*) form a quadrant. This thesis investigates the specific quadrant of *automated evaluation* of *model-generated text*.

I will start by discussing the qualities of human-written text and how humans assess them, and then list existing attempts to automate the evaluation process for human-generated text. I then move on to give an overview of natural language generation tasks, datasets and models proposed in recent years. This is to lay a broad foundation for the three generation tasks that this thesis is most concerned with, each of which will be discussed in detail in chapter 4. In section 2.3, I present two types of evaluation methods that are widely used for evaluating model-generated text in current NLG research - automatic evaluation and human evaluation.

2.1 Assessment of human-written text

Despite this thesis's emphasis on the evaluation of model-generated text, it remains essential to begin by examining the evaluation of human-written text as a source of inspiration and guidance for evaluating model-generated text.

In linguistics research, language usage has long distinguished between a person's first language (L1) and their second language or the language that they are currently studying (L2). The analysis of human-written texts is also closely associated with this context. In this section, I will discuss what is considered as a *text*, the aspects that denote the quality of a text, and how human assessment is usually carried out for text quality analysis. I will then explore existing attempts to automate the assessment process using computational

methods.

2.1.1 Text quality and human evaluation

Text quality is a broad topic. It is extremely difficult, if not impossible, to cover all aspects of text quality in a single chapter. What I attempt to achieve here is not a clear-cut, comprehensive definition of text quality and its constituent factors. Instead, my aim is to survey existing research in text quality and provide a supplementary background for the GFD evaluation framework, which is the core contribution of this thesis to the field.

In computational linguistics research, we talk about *text* all the time. What exactly does this terminology define? In discourse analysis, *text* is typically referred to as a cohesive structure of syntactic units such as words and clauses, whereas a non-text "consists of random sequences of linguistic units such as sentences, paragraphs or sections in any temporal and/or spatial extension" (Werlich, 1976). Hatim and Mason (2005) refers to *text* as "a set of mutually relevant communicative functions", and it is specifically structured "to achieve an overall rhetorical purpose". Though linguists have been actively arguing about the exact features that make a written passage a *text* for decades, most agree on the communicative function of text in human interaction. This purpose-driven definition places emphasis on the communicative aspect of a text. De Beaugrande and Dressler (1981) further claims that there are seven standards that a text has to meet to be considered as a communicative occurrence - cohesion, coherence, intentionality, acceptability, informativity, situationality and intertextuality. Without any of those standards, a text is not considered to be communicative. Non-communicative texts are considered as *non-texts* (De Beaugrande and Dressler, 1981). This indicates that there are certain rules that govern the reasons for a passage being a *text* or a *non-text*.

When composing a text, a writer needs to make linguistic decisions with regards to a wide range of aspects, including lexical choices, structural realisation of sentences, the overarching theme and the relationship between individual text segments and that theme. All these factors combine as linguistic indicators of *text quality*. The mastery of such indicators suggests a felicitous arrangement of text and so indicates an advanced level of writing competence.

As has been pointed out by Crossley (2020), the majority of work in writing quality assessment has been focusing on L2 learner texts. For L1 texts, the focus of research has mostly been on the study of *writing development* from children to adults (Aparici et al., 2021).

Text quality assessment for L2 English learners is typically performed using language ability tests (e.g., IELTS and TOEFL tests). In the writing section of these tests, L2 learners are asked to write a short essay in response to some test prompts that elicit free-text answers. The assessment of such essays aims not only to examine a learner's linguistic

knowledge such as vocabulary and grammar, but also to test a learner's higher-order cognitive skills regarding idea organisation, analytical thinking and logical reasoning. The assessment of L2 writing competence currently relies heavily on the judgement of human examiners. Consistent marking criteria are crucial for a fair scoring procedure. Take the second writing task in the IELTS test as an example. The official marking guideline[1] asks IELTS examiners to assess a student's writing quality based on the following criteria:

- **Task achievement** examines if an answer fully addresses the task with well-supported arguments.
- **Coherence and cohesion** examines a learner's ability to skilfully manage paragraphs in a cohesive way that it attracts no attention.
- **Lexical resource** assesses if there is a wide range of vocabulary and correct control of lexical features in the written text.
- **Grammatical range and accuracy** examines whether a learner is able to use a wide range of grammatical structures in their answer.

There are four aspects that are highlighted in the marking criteria. *Task achievement* posits a high-level requirement for an essay to closely address the question displayed in the task specifics. This examines the relevance of an answer to the writing task itself. *Lexical resource* and *grammatical range and accuracy* investigate important synthetic properties of a written passage. *Coherence and cohesion* is a more complex concept in that it implies not only *syntactic* attributes of a text being coherent in terms of sentence structure and lexical choice, but also *semantic* attributes such as thematic consistency and appropriate topical transition. However, the ways that these criteria are explained in the marking guidelines are not directly translatable to computationally executable schemes. For example, the guidelines state that an essay with good coherence and cohesion should attract "no attention" from the examiner. Although this is straightforward to human markers to comprehend, it is hard to formulate this requirement in a way that a computational model can be designed accordingly. Nonetheless, these guidelines continue to offer valuable insights for shaping the assessment of text quality for model-generated synthetic texts. What we appreciate and seek in human-written texts should similarly apply to the context of model-generated texts. The GFD evaluation framework presented in this thesis (discussed in detail in chapter 3) aligns with the aforementioned IELTS assessment criteria in several ways. The *grammaticality* measure in the GFD framework assesses *grammatical accuracy*, the *faithfulness* measure examines whether a model-generated output faithfully accomplishes the target tasks (i.e., *task achievement*), and the *diversity*

[1]See https://www.ielts.org/for-organisations/ielts-scoring-in-detail for detailed scoring guidelines.

measure looks into the extent of lexical and structural variety in the texts (i.e., *lexical resource* and *grammatical range*).

The IELTS scoring scheme is also echoed in modern linguistic studies. For instance, Dang (2006) proposed five dimensions to examine text quality: grammaticality, non-redundancy, referential clarity, focus, and structure and coherence. The underlying hypothesis of these evaluation proposals is that certain language features contained in a text can be used to predict the quality of the text.

category	**indicator**
lexical features	lexical diversity lexical density lexical sophistication
structural features	granularity (clause, sentence, T-unit) measure (frequency, ratio, index)
inter-text features	text coherence text cohesion

Table 2.1: Linguistic indicators for text quality used in existing literature.

In table 2.1 I summarise some of the linguistic features that have been used as indicators of text quality in conventional linguistics and computational linguistics studies. The three text quality factors - grammaticality, faithfulness and diversity - have been deliberately omitted from the table, as they will be thoroughly discussed in chapter 3.

The linguistic indicators of text quality can be broadly classified into three categories: *lexical features*, *structural features*, and *inter-text features*.

Lexical features examine straightforward linguistic clues such as lexical diversity (the ratio of *unique* words in a text), lexical density (the number of *content to function* words) and lexical sophistication (the ratio of *advanced* words in a text). Read (2000) suggests that the existence of sophisticated lexical units provides the richest information concerning text quality. Lexically sophisticated words are usually considered to be words that are less concrete and familiar, words that are more likely to be found in academic text, and words that have higher levels of latency in word naming and lexical decision tasks (Read, 2000; Kyle and Crossley, 2015, 2016; Laufer and Nation, 1995).

Synthetic features concern the level of sophistication and the variety of the synthetic forms in a given text. Synthetic measures look at three levels of granularity: clauses, sentences and T-units[2] (Hunt, 1965). For each level of granularity, the three primary methods for calculating synthetic complexity are frequency (the occurrence of a specific language unit), ratio (the number of a type of unit divided by the total number of all units or the number of another unit) and index (numeric scores computed by specific formulae)

[2]A T-unit is a minimal terminable unit of language, which refers to a main clause plus any subordinate clauses that may be attached to it.

(Wolfe-Quintero et al., 1998). The motivation behind using this class of linguistic indicator for quality assessment is straightforward. A skilful arrangement of these structural features is indicative of a great knowledge of the English grammar and a flexible application of a variety of language structures.

Inter-text features concern the connectivity of text segments within a given text, typically in the form of *text cohesion* and *text coherence*. A text with good coherence level makes it easier to maintain a reader's consistency of mental representation when reading the text. *Coherence* and *cohesion* are connected but subtly different concepts. *Cohesion* refers to the use of explicit linguistic cohesive devices, such as discourse markers and anaphora (Halliday and Hasan, 1976). *Coherence* is a broader concept in that it can be observed locally in terms of transitions between adjacent textual units, or globally in terms of overall topical coherence of the entire text. Cohesion can be roughly considered as an overt mechanism of discourse coherence that makes use of explicit linguistic cues. McNamara and Kintsch (1996) argue that, in contrast to cohesion which is treated as text-based, coherence is reader-based as it relies on individual understanding of a discourse derived from individual experiences and language proficiency.

In addition to the writer-focused qualities presented above, there are other metrics analysing text quality from a reader-focused perspective. An example of this is *readability*. Text readability is formally defined as "the sum of all elements in textual material that affect a reader's understanding, reading speed, and level of interest in the material" (Dale and Chall, 1949). Text readability is a reader-focused assessment as it examines text quality by analysing the relations between a given text and the efforts that readers make to understand it.

Different aspects of text quality are not entirely independent from one another. Certain aspects are intertwined and correlated in very nuanced ways. For example, lexical sophistication and text readability are two desirable criteria in writing. Lexical sophistication indicates a writer's capability of using advanced vocabulary. Readability describes the favourable feature of a text being accessible for a target group of readers. However, when the lexical sophistication of a text increases above a certain level, it requires significant efforts to comprehend the passage, therefore leading to a lower level of readability.

2.1.2 Attempts at automated evaluation of human-written texts

In section 2.1.1, I introduced the concept of text quality and the linguistic aspects that have been investigated when assessing human-written texts. This section discusses computational methods that have been proposed to automate the evaluation of text quality.

The evaluation of text quality is non-trivial. There has been an extensive amount of work on the automated assessment of text quality, which can be traced back to the 1960s (Page, 1966). Previous work in automated assessment (AA) of learner text has treated

text quality assessment as a supervised text classification or regression task. The goal of such a task is to assign a grade that indicates a learner's language level (Briscoe et al., 2010; Yannakoudakis et al., 2011).

Early studies in automated assessment use explicit clues in texts that can be highly predictive of specific attainment levels of a writer (Page, 1966; Landauer et al., 2003; Elliot, 2003). This is based on the assumption that lexical and grammatical properties are highly discriminative for linguistic competence revealed in a text. Briscoe et al. (2010) used a set of linguistic features, including lexical n-grams, part-of-speech n-grams, features representing phrase structure rules, and features representing other factors such as script length and error rate. Yannakoudakis et al. (2011) built on the above work by adding features that capture the syntactic complexity of sentences.

Aside from work that assigns a holistic grade for the overall text quality, many other studies have approached this research topic through focusing on specific attributes that reflect text quality, or inspecting text quality in specific use cases.

To measure **cohesion** between adjacent text segments, explicit cohesive devices are commonly used. One such device is a discourse marker (Halliday and Hasan, 1976) which signals text transitions. Other explicit linguistic cues include repeating lexical items in successive segments, or referring to previous text units using pronouns. Zhu and Bhat (2020) proposed GRUEN, which measured **coherence** by predicting a sentence-order prediction loss with a LITE BERT (Lan et al., 2019). Other works in the field of coherence assessment formulate the evaluation process as a classification task (Barzilay and Lapata, 2008; Elsner and Charniak, 2008; Farag et al., 2020). The implementation of such classifiers relies on artificial incoherent text. To construct negative samples, natural occurring texts are randomly permuted to serve as "incoherent" examples. The classifiers are then trained with the combined set of naturally occurring examples and artificially constructed incoherent examples. Xia et al. (2016) explored the research question of automated text **readability** assessment for L2 learners by analysing lexical and syntactic complexity, level of conceptual familiarity and logical sophistication. In Zhu and Bhat (2020), **non-redundancy** was evaluated by detecting repeated language units through explicit syntactic features. They also proposed an evaluation for text **focus** by examining the semantic consistency of adjacent sentences using word mover's distance (Kusner et al., 2015).

2.2 NLG tasks and models

The emphasis of this thesis is on the evaluation of text quality of *model-generated text*. As such, it is necessary to discuss natural language generation (NLG), which is commonly described as the construction of automated computer systems that can generate

understandable texts in a human language (Reiter and Dale, 1997).

This section provides a broad overview of widely studied tasks in the field of NLG, with the hope that the following discussions draw an overall landscape of existing NLG research. This will facilitate the in-depth exploration of the three tasks examined in the thesis (i.e., image captioning, data-to-text generation, dialogue generation) which will be discussed in chapter 4.

2.2.1 NLG tasks and datasets

NLG has existed since at least the development of ELIZA (Weizenbaum, 1966) in the 1960s. The most conventional notion of NLG involves generating natural language descriptions from an underlying representation of information. As Dale and Mellish (1998) put it, the underlying representation may be symbolic or numeric, but is generally non-linguistic. Typical examples of NLG tasks in this context include generating weather forecasts (Goldberg et al., 1994; Sripada et al., 2014; Reiter et al., 2005; Belz, 2007; Murakami et al., 2021) and automated journalism covering financial news (Kondadadi et al., 2013; Yan, 2022). Certain generation tasks, such as machine translation and summarisation, may not have been traditionally categorised as NLG tasks based on the non-linguistic data-to-text definition (Reiter and Dale, 2000). These fields have long been individually investigated as separate fields from NLG.

As the field has progressed, text-to-text generation tasks such as question answering and dialogue generation are now commonly referred to as NLG tasks as well. This unconventional usage of the term NLG encompasses a broader range of language generation tasks, covering image captioning, data-to-text generation, dialogue generation, and more. In this context, the initial constraint of the input being non-linguistic is not always strictly enforced.

These nuances in terminology usage represent customary practices within the NLP community. This thesis does not seek to establish any specific terminology concerning the precise scope of what constitutes NLG. However, it is important to include a disclaimer to clarify that in the subsequent sections, the term NLG is used in a broader sense, where the output consistently takes the form of textual language, while the input modality can be either textual or non-textual. In this context, current NLG tasks can be divided into two broad categories: *textual NLG* and *multimodal NLG*, based on the modality of the input source.

2.2.1.1 Textual NLG

Textual NLG is a form of NLG where both the input and the output of a system are in a textual format. In the following I will briefly introduce popular sub-tasks and

accompanying datasets in textual NLG, including machine translation, text summarisation, question answering (QA) and dialogue generation.

Machine translation is a classic task in NLG which can be traced back to the 1950s (Hutchins, 2000). Given a sentence in a source language, the task of machine translation aims to automatically translate the sentence into a target language. Machine translation has made rapid progress in recent years, especially with the popularisation of WMT shared tasks (Bojar et al., 2014, 2016, 2017; Kocmi et al., 2022) and easily accessible large-scale parallel datasets in various language pairs (Lison and Tiedemann, 2016; Koehn et al., 2003). Alongside the conventional *source language - target language* translation framework, existing literature has also inspected particular areas in machine translation. For example, the Machine Translation of Noisy Text dataset (Michel and Neubig, 2018, MTNT) investigated the translation process when the training corpus contains naturally occurring noisy texts. Datasets with a specific focus on automatic post-editing (Fomicheva et al., 2022; Góis et al., 2020; Negri et al., 2018, APE) explore the automated post-editing effort of correcting machine translation errors to obtain higher-quality translations. For long-tailed low-resource languages, the lack of publicly available parallel corpora has been a major issue in machine translation for these languages. Researchers in this field have been endeavouring to address this issue. OPUS-100 (Zhang et al., 2020a) is a such attempt where the authors collected a multilingual dataset with 100 languages and explored machine translation approaches when one end of the translation is a low-resource language. In a similar work, Goyal et al. (2022) collected the FLORES-101 dataset to enable better assessment of translation quality on low-resource languages.

Text summarisation in NLG is the process of summarising large pieces of texts into short, accurate summaries whilst keeping as much vital information as possible in the summarised outputs. There is an enormous amount of literature introducing datasets curated for the particular task of text summarisation. These datasets can be roughly grouped based on the type of text sources that are summarised, including news articles (e.g., CNN/DailyMail (Nallapati et al., 2016), Multi-News (Fabbri et al., 2019), NYT Corpus (Sandhaus, 2008), XSum (Narayan et al., 2018), DUC (Copeck et al., 2006, 2007)), academic papers (e.g., S2ORC (Lo et al., 2020)), dialogues (e.g., SAMSum (Gliwa et al., 2019)), Wikipedia pages (e.g., WikiHow (Koupaee and Wang, 2018)), books and stories (e.g., BookSum (Kryściński et al., 2021)), and Reddit posts (e.g., Reddit TIFU (Kim et al., 2019), TLDR-17 (Völske et al., 2017)).

Question answering (QA) is the task of automatically generating an answer to a question given some context or knowledge. Most QA datasets originate from the task of reading comprehension where a question is raised inquiring about a specific detail in a given passage. In this case, the generation of an answer is significantly grounded by the content of the passage. The collection process of such reading comprehension

datasets typically use crowd-sourcing approaches. When presented with a piece of article or document, crowd workers are asked to create questions regarding the contents in the given text, and then produce correct answers to those questions. The given text passage can be in the form of news articles (e.g., NewsQA (Trischler et al., 2017)), social media texts (e.g., TweetQA (Xiong et al., 2019)), Wikipedia articles (e.g., QuAC (Choi et al., 2018), SQuAD 2.0 (Rajpurkar et al., 2018)), cooking recipes (e.g., RecipeQA (Yagcioglu et al., 2018)), books and movie scripts (e.g., NarrativeQA (Kočiskỳ et al., 2018), DuoRC (Saha et al., 2018)), knowledge bases (e.g., SimpleQuestions (Bordes et al., 2015) based on Free base, QALD-9 (Ngomo, 2018) based on DBpedia). The recently collected CoQA dataset (Reddy et al., 2019) provide QA pairs associated with text passages spanning over seven domains, including children's stories, literature, news, Wikipedia articles, Reddit posts, high school exams and scientific articles.

Another group of QA datasets explore general question answering. These datasets usually contain real-world QA pairs (often in the form of anonymised queries from search engine logs), augmented with accompanying text snippets (often Wikipedia pages retrieved by Google). The related text snippets may or may not provide enough information to answer the questions. An example of such datasets is WikiQA (Yang et al., 2015), which uses real-world Bing queries as the question source and each question is linked to a Wikipedia page that potentially contains the information needed to answer the question. Many other recent datasets were constructed along similar lines, such as Natural Questions (Kwiatkowski et al., 2019), MS MARCO (Nguyen et al., 2016) and SearchQA (Dunn et al., 2017).

Recently, a number of QA datasets have been constructed to examine specific aspects of textual question answering. DROP (Dua et al., 2019) is a challenging QA dataset which consists of complex, compositional questions which require discrete reasoning abilities (e.g., addition, counting, sorting) to answer. The bAbI task (Weston et al., 2015) was proposed to examine QA models' reasoning abilities over a chain of facts when predicting an answer. HotPotQA (Yang et al., 2018) is a multi-hop QA dataset in which a question is answered by inferring across multiple documents. TriviaQA (Joshi et al., 2017) is another complex QA benchmark which provides compositional questions with a higher level of syntactic variability, thus requiring reasoning over multiple sentences to generate an answer.

Dialogue generation has drawn much research interest in the past few years. Existing work in this field can be essentially classified into two groups with regards to their domain restrictions. Domain-specific task-oriented dialogue tasks (Budzianowski et al., 2018; Galley et al., 2019) require dialogue agents to help users complete pre-defined goals in specific domains. Open-domain dialogue tasks (Chen and Kan, 2013; Zhang et al., 2018a; Li et al., 2017; Tang et al., 2019) allow agents to have open-ended conversational interactions with users, typically in the form of online chit-chats (Shaikh et al., 2010; Forsyth and

Martell, 2007).

2.2.1.2 Multimodal NLG

Driven by mature technologies in the various fields of artificial intelligence and the release of large-scale datasets, there has been an increasing interest in combining different modalities in a single generation task.

Visual question answering (VQA) is a significant and technically challenging multimodal NLG task that connects language and vision (Antol et al., 2015). In the general form of VQA, the computer is presented with an image and a natural language question about this image. The computer is then asked to predict the correct answer to that question. A large number of real-world datasets have been released specifically for VQA. An early attempt to collect large-scale VQA dataset is the DAtaset for QUestion Answering on Real-world images (Malinowski and Fritz, 2014, DAQUAR). Question-answer pairs are collected in two different ways. The synthetic question-answer pairs are automatically generated using predefined templates. Another set of question-answer pairs are annotated by human participants for the images. The final dataset consists of 795 images accompanied by 6794 question-answer pairs for training, and 654 images with 5674 question-answer pairs for testing. Although DAQUAR enables the development of early neural VQA methods, its major disadvantage lies in its indoor setup, which limits the types of images and thus the questions that can be asked. Later work in VQA datasets (e.g., the VQA dataset (Antol et al., 2015), the COCO-QA dataset (Ren et al., 2015)) are built on the MS-COCO dataset (Lin et al., 2014) which contains a large number of complex, real-world scenes. Visual Genome (Krishna et al., 2017) was proposed to promote the cognitive modelling of objects, attributes and their pairwise relationships in visual scenes, with rich annotations for over 108K images.

Image captioning is another multimodal NLG task that has been rigorously studied. Image captioning aims to automatically generate descriptive captions for images, and has important practical applications; for example, when combined with speech technologies, it can help visually impaired people obtain a better understanding of the content of images. Automatically generating image captions requires computers to be able to understand images and generate descriptive sentences according to the feature representations learnt from the images. Although VQA and image captioning both involve visual and linguistic understanding, VQA focuses more on specific details whilst the general form of image captioning requires a high-level grasp of visual content. A large number of datasets have been proposed for the task of image captioning, including MS-COCO (Lin et al., 2014), Flickr8k (Hodosh et al., 2013), Flickr30k (Young et al., 2014), VizWiz-Captions (Gurari et al., 2020) and SentiCap (Mathews et al., 2016).

Audio visual scene-aware dialogue (AVSD) was proposed in one of the DSTC

tracks[3] (Hori et al., 2018; Alamri et al., 2019; Hori et al., 2019). Given a video clip along with its soundtrack, AVSD aims to generate a response to a user input question in the context of a given dialogue. Textual summaries of the videos are also provided. The AVSD task extended the VQA task with increased complexity in both the linguistic and the visual modalities. Instead of using images, one of their main input sources consists of video clips selected from the Charades dataset (Sigurdsson et al., 2016), which consists of short videos of daily household activities. The generation of a textual answer is grounded by multiple modalities, including vision (video clips), audio (soundtracks from the videos) and text (dialogue history, the question to be answered, and video summarises).

Existing multimodal datasets often contain a large number of training and testing instances, making them well-suited for training and evaluating deep neural networks. However, these kinds of crowd-sourced multimodal datasets are notoriously prone to undesirable biases. For example, multimodal models have been observed to be able to achieve good results on various benchmarks (Hori et al., 2018; Antol et al., 2015) by only relying on language priors to predict outputs, ignoring to a significant extent other modalities that may provide useful information (Zhang et al., 2016b; Agrawal et al., 2016; Goyal et al., 2017). In response to this, studies have explored a wide range of approaches to identify the unfavourable biases contained in certain datasets (Agrawal et al., 2018; Kervadec et al., 2021; Dancette et al., 2021) and to reduce dataset biases (Zhang et al., 2016b; Cadene et al., 2019).

2.2.2 NLG models

NLG techniques range from simple rule-based systems which generate text using pre-defined templates, to considerably more complex statistical systems which use large corpora of human-written texts. There are survey papers (Gatt and Krahmer, 2018; Reiter and Dale, 2000; Dong et al., 2022; Erdem et al., 2022) which give a detailed account of the history of NLG models. A comprehensive depiction of existing NLG models is beyond the scope of this thesis. For reasons of space I will only cover models that are relevant to this thesis.

Rule-based models. NLG systems traditionally used hand-crafted templates or rules written by domain experts (Binsted and Ritchie, 1994; Reiter et al., 1995; Osman et al., 1994; Robin and McKeown, 1996; Dale, 1990; Reiter et al., 2003; Goldberg et al., 1994). Reiter and Dale (2000) pipelined NLG frameworks into three stages: document planning, microplanning and surface realisation. Rule-based systems are suitable for cases where the application domain contains a limited number of patterns and the rule-set is straightforward to construct. Building a rule-based system becomes more difficult when the domain contains complex narratives and rules need to be constructed for diverse edge cases. Due to the necessity of deploying extensive expert knowledge, rule-based approaches

[3]http://workshop.colips.org/dstc7/call.html

tend to incur high maintenance costs, therefore limiting their scalability to new patterns and their generalisation to new domains.

Conventional machine learning models. The access to suitable corpora has enabled the transition from rule-based approaches to statistical approaches. The term *conventional machine learning* here refers to the data-driven, statistical approaches that were widely used before the popularisation of deep neural networks. These machine learning approaches contrast with rule-based approaches in that they do not require hand-coded rules for generation. Instead, they rely on carefully selected features and representative data samples to learn a predictive function based on the data. There have been many successful attempts in applying data-driven techniques to the field of NLG. Examples include the BAGEL system (Mairesse et al., 2010) which uses dynamic Bayesian networks to generate phrases from linearised semantic trees. In related work, Konstas and Lapata (2012) investigated a data-driven method to generate textual descriptions from database concepts. A classic application in machine learning NLG is statistical machine translation (SMT) (Chiang, 2005; Koehn et al., 2007; Koen, 2004), in which alignments of source and target languages are explicitly learnt from bilingual corpora. Before the introduction of neural translation models, SMT was the most widely used translation method. SMT models were initially word-based (e.g., IBM HMM models (Brown et al., 1993; Vogel et al., 1996; Gal and Blunsom, 2013), GIZA++ (Casacuberta and Vidal, 2007)), superseded by phrase-based translation models where the fundamental unit of translation is a phraseme (different from linguistic phrases) found in parallel corpora. Facilitated by the great success of SMT systems, Manishina (2016) proposed a method to cast other NLG tasks as a kind of translation task so that SMT can be used for such generations. Their attempt showed decent performance, comparable to that of template-based generation systems. Statistical methods have the merit of automatically learning a desired representation of real-world data points which can be used to predict unseen input, thus liberating researchers from manual template building. However, their heavy dependence on annotated data shifted human effort from template construction to dataset creation and feature engineering, which appears to be another instance of the "no free lunch" principle in machine learning.

Neural networks. The advance of text vectorisation approaches (e.g., Word2Vec (Mikolov et al., 2013), GLoVe (Pennington et al., 2014)) promotes the use of neural networks in language-related tasks. The vectorisation process maps language tokens to numeric vectors whilst retaining some level of syntactic and semantic attributes. The resulting vectors are suitable for use as an input to neural networks. The above vectorisation methods are generally referred to as *static* approaches, as their mapping schemes are usually pre-trained and kept frozen. Hence they are notoriously bad at handling out-of-vocabulary (OOV) words and context-dependent words. Before the advent of Transformer-based models (Vaswani et al., 2017), recurrent neural networks (Elman, 1990, RNNs) were

arguably the most widely used model structure in NLP research. In practice, gated recurrent units (Cho et al., 2014, GRUs) and long short-term memory models (Hochreiter and Schmidhuber, 1997, LSTMs) are often used as improved variants of the recurrent neural architecture. Built upon RNNs, a sequence-to-sequence (seq2seq) encoder-decoder framework (Sutskever et al., 2014) was introduced. This model architecture was first tried out with machine translation, but has since gained huge popularity in other NLG tasks. The encoder and the decoder of a seq2seq model are both RNNs. The encoder traverses the input and embeds the input information into a hidden representation, from which the decoder gradually outputs a sentence by predicting one word at a step until the end-of-sentence (EOS) special token is generated.

The recently proposed Transformer model (Vaswani et al., 2017) is based on the encoder-decoder framework, with a carefully designed multi-head self-attention mechanism in its encoder. This model architecture has increasingly become the model of choice for NLP tasks and has achieved state-of-the-art performance on a wide range of benchmarks (So et al., 2019; Dai et al., 2019; Rae et al., 2019; Yan et al., 2019) since its appearance. Inspection of the Transformer model and potential improved variants is still an active research field. The training parallelisation in Transformer facilitates training on larger-scale data, which in turn led to the development of contextual pre-trained models such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018). The BERT model is a language model pre-trained with multi-head Transformer encoders. The training process of BERT makes use of the context from both directions of an input token (i.e., bidirectional), allowing it to capture contextual information from both preceding and following words. BERT is trained using a masked language modelling (MLM) objective, where a certain percentage of words in the input are randomly masked. The model is trained to predict those masked words based on the context provided by the other words. BERT is typically used for text classification but has been also adapted to generation tasks (Rothe et al., 2020). On the other hand, the GPT model is a unidirectional Transformer, trained with an auto-regressive language modelling objective where the model learns to predict the next word in a sequence given the previous words. The model is trained to maximise the likelihood of the target word given the context. GPT is usually considered to be suitable for language generation tasks. These two models are both pre-trained with large datasets and allow further fine-tuning of word embeddings in particular tasks. Other pre-trained NLG models have also achieved impressive results on various benchmarks, such as the BART model (Lewis et al., 2019) which adopts a bidirectional encoder (same as BERT) and a unidirectional decoder (like GPT), the T5 model (Raffel et al., 2020) which is an encoder-decoder language model pre-trained on a mixture of supervised and unlabelled data and converts every language task into a text-to-text format, and the now much discussed ChatGPT model (OpenAI, 2022) which is first trained using supervised

fine-tuning and is then further optimised using Reinforcement Learning from Human Feedback (Knox and Stone, 2008; Christiano et al., 2017, RLHF).

Compared to conventional machine learning approaches, the end-to-end property of most neural architectures relieves developers of the previously-necessary endeavour of manual feature engineering. However, researchers have expressed concerns about the lack of interpretability and the heavy dependence on hyperparameter tuning of such models. The fickleness of these models makes it harder to obtain consistent and reliable understanding of model behaviours. As van Miltenburg et al. (2020) pointed out, we have seen "the move away from rule-based systems as a means to evade responsibility for whatever output our NLG systems produce", as the learning process for statistical models is greatly influenced by the texts crowd workers annotated in the collected corpora. More and more researchers have now realised that, by relieving ourselves from the extensive labour of feature engineering, we have exempted ourselves from the responsibility of controlling the generation behaviours of our models, especially for complex generation scenarios and edge cases. For this reason, there has been a recent swing-back to a rule-based *perspective* which urges more engagement with the datasets that we construct, and a more controlled generation processes (van Miltenburg et al., 2020).

Multimodal models. The typical framework for multimodal NLG tasks (Perez et al., 2018; Ben-Younes et al., 2017) is to extract features for individual modalities separately, and to then fuse them into a single representation, from which an output sequence can be decoded. Convolutional neural networks (Krizhevsky et al., 2012; Deng et al., 2009, CNNs) are often used to extract visual representations. Recurrent neural networks (RNNs) and their variants (Elman, 1990; Hochreiter and Schmidhuber, 1997; Graves et al., 2013; Peris and Casacuberta, 2015) are used to either encode textual information in the input modalities, or decode a fused multimodal representation to a natural language sentence. Recently, vision transformers (Dosovitskiy et al., 2021, ViT) have been proposed specifically for multimodal tasks. Swin Transformer (Liu et al., 2021c) is a hierarchical vision transformer which computes its representation via shifted windows, thus limiting self-attention computation to "non-overlapping local windows" and allowing image modelling at various scales. Another vision transformer - TimeSformer (Bertasius et al., 2021) - adapts the vision Transformer framework to video by extending the visual self-attention mechanism to allow the coverage of space and time. The release of these pre-trained multimodal Transformers greatly facilitates research in the field of multimodal tasks.

2.3 Evaluation of model-generated text

Evaluation metrics are crucial to the R&D of NLG systems, as they serve as a proxy for performance in real applications. Precisely quantifying the performance of NLG systems has been a challenging problem, as judging the quality of candidate outputs is ultimately a context-dependent and subjective matter.

NLG models have long been tested along two broad lines: *extrinsic* evaluation and *intrinsic* evaluation (Belz and Reiter, 2006). *Extrinsic* evaluation assesses a model's performance by integrating the model into a real-world task or pipeline that it is designed for, and evaluating how well the model improves the overall performance of the task or pipeline. For example, to measure the effectiveness of an NLG system that produces personalised smoking-cessation letters, Reiter et al. (2003) recruited several thousand smokers and divided them into two groups: an experimental group that received personalised letters and a control group that received fixed non-personalised letters. They collected the number of people who quit smoking after they received the letters in each group, and measured whether the personally tailored letters changed people's smoking behaviours as the researchers had hoped. As such, extrinsic evaluation highlights the ultimate aim of NLG systems to be helpful for real-world tasks. Whereas extrinsic evaluation focuses on whether a model achieves its desired outcomes in real-world context, *intrinsic* evaluation directly examines the model's outputs. There has been a significant decrease in the number of NLG papers that use extrinsic evaluation. van der Lee et al. (2019) hypothesised that as the focus of NLG research has recently shifted from pipelines (e.g., Reiter and Dale (1997)) to sub-tasks of the traditional pipelines, intrinsic evaluation itself may be sufficient for many of the recent studies. Moreover, extrinsic evaluation can be expensive and time-intensive as the final judgement of performance often requires expert knowledge. Potentially due to these factors, the use of extrinsic evaluation has become increasingly rare in contemporary NLG papers. It should be emphasised that the evaluation approaches discussed in the rest of the thesis refer to **intrinsic** evaluation, unless otherwise stated.

2.3.1 Human evaluation

Human evaluation is generally considered as the gold standard for NLG evaluation. It is seen as the most reliable indicator of progress in the NLG field. Despite wide usage, many challenges remain in planning and reporting human evaluation. The practice of human evaluation suffers from a variety of issues including the usage of heterogeneous (and therefore hard to compare) evaluation processes, lack of standard protocols, and missing details from reported results (such as the number of annotators involved).

When evaluating NLG systems for weather forecasts, Belz and Reiter (2006) found that the judgement from an arbitrary non-expert annotator does not correlate well with

average expert judgements. From this we can see the impact of the expertise (or lack of) of participating annotators on evaluation credibility. Läubli et al. (2020) investigated how the design of human evaluation affected the results achieved by machine translation systems, and discovered that the perceived text quality depends heavily on "the choice of raters, the availability of linguistic context, and the creation of reference translations". They also found that inter-annotator agreement was more consistent between expert raters than non-experts. In addition, expert raters were able to identify a larger gap between human-written sentences and model translations due to their higher sensitivity to translation nuances. Recently, van der Lee et al. (2019) provided a broad overview of recent developments and suggested practical suggestions for conducting better human evaluation. The authors examined papers from ACL and INLG in 2018, and found that there is currently no consensus as to how NLG systems should be evaluated. For example, different studies employ different number of annotators when carrying out human evaluation, and only 12.5% of the papers reported their inter-annotator agreement scores. There is also a lack of agreement on the number of candidate sentences that are evaluated in an evaluation process. Similarly, Howcroft et al. (2020) evaluated 20 years worth of NLG papers (from 2000 to 2019) that reported human evaluation results from two NLG-specific conferences, INLG and ENLG. They found that there are more than 200 different terms that have been used to describe evaluated aspects of language quality. The terms used in those papers are themselves poorly defined. Moreover, the evaluation results seem to be easily perturbable if using a different selection of evaluation metrics. As a result, it is hard to either replicate the human evaluation results of published papers or compare experiment results across papers.

There is a wide acknowledgement of the importance and necessity of high-quality human evaluation in the NLG community. NLG researchers stand to benefit from a clearer protocol of human evaluation practice and a sounder method of result reporting. Based on empirical findings, Läubli et al. (2020) offered a set of recommendations for those evaluating strong machine translation models in terms of human-machine parity. These recommendations encouraged researchers to require professional translation expertise from raters, and suggested providing access to original source texts and full document-level context for the raters. Howcroft et al. (2020) provided a set of normalised quality criterion names as a starting point for a standardised evaluation set. They also advocated that the reporting of evaluation results should be arranged with great caution, covering considerations regarding the exact instructions that were given to the raters in an evaluation, the detailed definitions of all quality criteria that are measured, and the details of the rating approach (e.g., ranking vs rating scales). In a similar vein, van der Lee et al. (2021) highlighted the requirement for a clear definition of what is being evaluated and the goal of the evaluation before carrying out any human evaluation. They recommended using

multi-item 7-point Likert scales or continuous ranking as the rating method. Furthermore, they emphasised the (currently often neglected) necessity of making the materials related to the evaluation process publicly available to assist in future replication.

2.3.2 Automatic evaluation

Evaluating text quality by human preference is slow and laborious. To alleviate this problem, many automatic evaluation metrics have been proposed. In the following, I will introduce the automatic evaluation methods that are widely used across NLG tasks, followed by metrics that are specifically proposed for particular NLG tasks.

N-gram based metrics. N-gram based metrics are arguably the most commonly used automatic evaluation metrics in NLG fields. The rationale of using these evaluation metrics is that human-written references serve as the approximate target of generation, and comparing model outputs to this target on the basis of n-grams is taken as a proxy for how well a system performs. In this way, a model-generated candidate is not directly evaluated with respect to the restrictions set by the input or task requirements, but compared to a set of human-generated statements.

N-gram based metrics approximate the quality of a candidate output by comparing the surface n-gram similarity of the candidate with a number of human-generated references. BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015) and LEPOR (Han et al., 2012) all fall under this category. BLEU calculates the weighted geometric mean of overlapping n-gram scores with an extra brevity penalty for short sequences, whilst METEOR further takes into account synonyms and phrase matching to capture semantic similarity. CIDEr extends existing n-gram-based methods with word stemming and *tf-idf* weighting, and LEPOR takes into account the factors of n-gram word order penalty, precision, recall and length penalty. These measures were first proposed for the evaluation of machine translation systems. They have since been applied in a range of settings for NLG evaluation. Similarly, ROUGE (Lin, 2004) is an n-gram based metric which was initially proposed for text summarisation evaluation, and has been widely deployed for the evaluation of other NLG tasks. Compared to other n-gram metrics, ROUGE presents a different emphasis on n-gram overlapping, skip-bigram based co-occurrence and longest common subsequence (LCS) matching.

Investigating improved n-gram based evaluation metrics is an active research field in NLG. There have been a number of variations proposed for the above metrics. For example, NIST (Doddington, 2002) extended BLEU by adding different weights to particular n-grams based on their informativeness. That is to say, the rarer an n-gram is in a corpus, the more weight it will be given by NIST.

Despite their huge popularity, many researchers have noted issues that manifest when using such metrics for NLG evaluation. Elliott and Keller (2014) reported *weak* to *moderate*

correlations between n-gram based metrics (i.e., BLEU-1, ROUGE, METEOR) and human judgement when evaluating image captioning outputs. This has been repeatedly reported in many other studies (Elliott and Keller, 2014; Anderson et al., 2016; Kilickaya et al., 2017). When investigating evaluation metrics for a text summarisation task, Moramarco et al. (2022) discovered that the choice of human references has a significant effect on all n-gram based metrics. The main problem with n-gram based metrics is their heavy dependence on surface n-grams in a fixed set of references. These metrics treat a set of references as a proxy for the desired generation target, and proceed from this idealised abstraction. Fundamental concerns have been raised with respect to the validity of BLEU as a metric for tasks other than machine translation in general (Reiter, 2018), particularly for tasks for which the output content is not narrowly constrained (Liu et al., 2016). Furthermore, there have been questions regarding the variability of such metrics in parameterisation leading to significantly different results (Post, 2018).

Embedding based metrics. In contrast to n-gram metrics which are generally believed to predominantly capture surface syntactic similarity, embedding based metrics aim to evaluate model outputs by comparing their meanings to those of the references. This is achieved through the use of contextual embeddings derived from word embedding models (e.g., Word2Vec (Mikolov et al., 2013), ELMo (Peters et al., 2018)). Embedding based evaluation metrics map words into their vector representations. Similarity between a model output and a target reference is computed by measuring the similarity of their embeddings. There have been many automatic metrics proposed on this basis. In BERTScore (Zhang et al., 2019b), a ground truth reference and a model-generated output are both passed through a pre-trained BERT model to generate contextual word embeddings. These two sets of word embeddings are greedily matched (Corley and Mihalcea, 2005; Rus and Lintean, 2012), from which a number of measures (i.e., precision, recall, F-score) are calculated. Instead of greedy matching, MoverScore (Zhao et al., 2019) uses optimal matching for word embeddings based on Word Mover's Distance (Kusner et al., 2015). Other similar works include ROUGE-WE (Ng and Abrecht, 2015), MEANT 2.0 (Lo, 2017), YiSi (Lo, 2019), BLEURT (Sellam et al., 2020), among others.

Supervised metrics. Supervised metrics such as BLEND (Ma et al., 2017), RUSE (Shimanaka et al., 2018) and SUM-QE (Xenouleas et al., 2019) are trained to optimise their correlation with human evaluation. This training paradigm allows a desirable, direct approximation of human judgements. However, these metrics require large numbers of human annotations to train them, and are prone to generalisation issues when being deployed in new domains.

The metrics presented in the list above are *general-purpose* evaluation approaches that were either initially designed or eventually used for evaluation across NLG tasks. There are also many metrics that are specifically proposed for *particular* tasks. For example, SPICE

(Anderson et al., 2016) is an automatic evaluation metric which makes use of semantic propositions in scene graphs for image captioning evaluation. SUPERT (Gao et al., 2020) proposed an unsupervised metric for text summarisation, which rates the quality of a summary by measuring its semantic similarity with selected sentences from the source document. For text simplification, Xu et al. (2016) designed SARI, an evaluation approach which compares model outputs not only to the references, but also to the input sentences.

Automatic evaluation serves as a quick and repeatable way to approximate text quality. But when aiming to achieve a general evaluation of overall quality, human evaluation remains the gold standard in the NLG field. Research into automatic evaluation seeks to find a **sufficient surrogate** for human evaluation. An ideal automatic evaluation would not only be quick to perform, but would also correlate well with human judgements. The validity of automatic evaluation metrics is usually evaluated via the assessment of their correlations with human preference. There are a variety of methods used in NLG practice to measure such bivariate correlations, which I briefly summarise below.

Pearson's correlation (Pearson, 1895) assesses the strength of linear association between two continuous variables. The calculation of Pearson's correlation depends on a number of assumptions. Both variables should follow a normal distribution, and the relationship between the variables should be linear. The Pearson's correlation coefficient always returns a value between -1 and 1. A value of 1 indicates a perfect positive linear relationship, and -1 indicates a perfect negative linear relationship. A value of 0 signifies no relationship between the two variables. In contrast to Pearson's which measures two variables' linear dependence, **Spearman's rank correlation** (Spearman, 1904) measures the monotonic relationship between two ranked variables. Spearman's is a non-parametric measure, as it does not require the two variables to fall into a bell curve. **Kendall's Tau correlation** (Kendall, 1938) is very similar to Spearman's and can often be used interchangeably with Spearman's. Like Spearman's correlation, Kendall's correlation also measures a monotonic relationship using ranked data.

Chapter 3

Three metrics: Grammaticality, faithfulness and diversity (GFD)

Some observers in the field of NLG have noted the presence of a negative cycle in which people report improvements based on under-verified evaluation methods that are not inherently suitable for such evaluation (Callison-Burch et al., 2006; Liu et al., 2016; Reiter, 2018). In response to this, the core of this thesis is to explore a principled evaluation framework for NLG tasks with an explicit focus on grammaticality, faithfulness and diversity (GFD). The theme of this thesis is not to propose a one-for-all paradigm for NLG evaluation. That is way beyond any realistically possible single piece of work. Instead, I aim to raise again the great importance of evaluation methods for NLG tasks and demonstrate how a principled, diagnostic evaluation framework can help identify model strengths and limitations.

This chapter commences with an explanation of the rationale for selecting GFD as the evaluation metrics, followed by a comprehensive examination of these metrics, which are individually presented in their respective sections.

3.1 Rationale for the choice of metrics

Many evaluation methods that are widely used in current NLG research (Papineni et al., 2002; Banerjee and Lavie, 2005; Lin, 2004; Sellam et al., 2020; Zhang et al., 2019b) report a holistic score as an indicator for model performance. Hastie and Belz (2014) argued that evaluating a text through its overall delivery of the communicative goal is often too abstract to computationally measure. As such, they advocated the use of separate criteria weighted according to their importance to the overall goal. In alignment with their work, I propose an evaluation framework for NLG tasks that prioritises three distinct aspects - grammaticality, faithfulness and diversity. These aspects correspond to fundamental requirements for NLG models: (a) that the outputs are grammatically well-formed, (b)

that the information contained in the outputs is consistent with the task prior and input contents, and (c) that the outputs exhibit diverse syntactic constructions. Among these three metrics, grammaticality and diversity assess the quality of sentences generated by models at the form level, whilst faithfulness assesses the quality of model outputs at the content level.

The above criteria are arguably fundamental requirements for any kind of text generation systems. The importance and necessity of such measurements have been repeatedly recounted in linguistic literature. The three measures also align well with the evaluation criteria established by many language assessments for texts written by L2 learners (refer to the previous section 2.1.1 for detailed discussion). The three metrics are carefully chosen to be as orthogonal as possible, so they measure relatively different things with regards to text quality. **Grammaticality** posits an inherent syntactic requirement for a sentence to be grammatically well-formed regardless of the concrete context. I consider this to be the most fundamental criterion for the evaluation of text quality for NLG model-generated outputs. **Faithfulness** assesses whether an output is compatible with the constraints placed by the target task itself, as well as the content of the input. Whilst grammaticality and faithfulness measure the degree of formal or contextual "correctness" of an model-generated output, **diversity** measures the degree of variability at the surface form, providing a complementary perspective for evaluation. By assessing diversity, the evaluation framework aims to capture the model's ability to generate outputs that are not only grammatically and semantically correct, but also exhibit variability in their form-level features. The concept of diversity at the form level pertains to explicit variations in the lexical and syntactic components within a given text. This is in contrast to implicit dimensions of diversity that relate to semantics and are usually more difficult to measure or quantify. The metric of diversity is inherently an inter-textual measure, as its applicability relies on the comparison of multiple texts and the respective complexity of those texts. In cases where only a single sentence is under consideration, defining the concept of diversity can be challenging, if not impossible. As a result, this thesis conceptualises diversity as the syntactic variability observed across the entire corpus or the group of model outputs, and can be evaluated at both the lexical and constructural levels. Specifically, word-level diversity assesses the variety of lexical resources, and construction-level diversity pertains to the grammatical variability present in the generated text. In this context, the terms "diversity" and "variability" are often used interchangeably.

By dividing the overarching goal of evaluation into three granular evaluation qualities, we are able to gain insightful information regarding model performance in distinct aspects. The three evaluation metrics, namely faithfulness, grammaticality, and diversity, serve distinct purposes. The faithfulness evaluation lays emphasis on content verification in text, whereas the grammaticality and diversity assessments focus on qualities related

to syntax. Grammaticality and faithfulness operate at the sentence level, scrutinising individual sentences for their adherence to syntax rules and content accuracy, respectively. In contrast, diversity is a corpus-level measure, examining the range of syntactic structures and word usage across an entire corpus of text. This layered approach of GFD evaluation serves as an evaluation protocol encapsulating the necessary facets of assessing NLG systems, enabling better understanding of the performance of NLG models across different dimensions of text generation.

It is worth pointing out that measures such as cohesion and coherence mentioned in section 2.1.1 are not considered in the GFD framework, as the target outputs evaluated in this thesis are usually short-form texts that consist of 1-3 sentences, most of them with single sentences. As a result, it is sensible to focus on sentence-level evaluations rather than metrics that usually look at inter-sentence transition (cohesion) or overarching consistence (coherence). In addition to this, text readability will not be considered in the present discussion given its status as a high-level criterion (comparable to that of "text quality"), which encompasses multiple more granular qualities. As reiterated throughout this thesis, my prime objective is to establish a fine-grained evaluation framework, with a focus on diagnosing critical aspects of model outputs. To achieve this, it is imperative to carefully select the criteria for analysis, as the chosen metrics need to satisfy the requirements of being both essential for the text generation scenario and amenable to computational formulation.

3.2 Grammaticality

I begin this section with an exploration of the definition of the term "grammaticality". Following this, I will present the ongoing debate concerning prescriptive and descriptive methods in linguistic research. I will then introduce the English Resource Grammar (ERG), the grammar underpinning the evaluation for grammaticality in this thesis, and discuss related studies on automated grammaticality evaluation.

3.2.1 What is grammaticality?

A text at the first sight might look like an arbitrary ensemble of lexical symbols. Even with a limited set of words, we are able to summon up a massive number of combinations of potentially acceptable expressions. It is widely recognised that humans do not generate incoherent collections of words without structure. Instead, we employ syntactic constructions that have been either explicitly or implicitly established within certain communities to facilitate the communication of ideas and information. It can be argued that grammaticality is a characteristic of language that is inherently embedded in human communication, whether it is strictly or loosely enforced.

Grammaticality typically refers to a binary concept that describes whether a sentence adheres to the rules of grammar for a particular language. While this definition may initially appear straightforward to understand, determining whether a given text is grammatical is a more complicated issue than it might seem at first glance. After all, the question of what constitutes grammaticality in a sentence is not a straightforward one.

When asked what exactly makes a sentence a valid instance of a language, even native speakers may struggle to come up with a definition. However, virtually every native speaker is capable of quickly gauging whether a sequence of words arranged in a particular manner forms a natural-sounding and acceptable sentence in their language. Prior to delving into types of grammars that have been used to determine grammaticality (discussed in section 3.2.2), I will first explore the existence of what is commonly referred to as "standard English", and the commonly used terminologies when referring to concepts related to grammaticality.

It has been argued that *standard* English is an artificial construct, as language is prone to variation and is known to constantly evolve over time (Bickerton, 1991; Aitchison, 2005; Locke and Bogin, 2006). Diverse cognitive and cultural factors can contribute to shifts in language use that occur over time. So-called standard English at the time of writing this thesis is unquestionably different from the English of twenty years ago, not to mention more distant times like Shakespeare's era. Additionally, the judgement of grammaticality at a particular time depends on the context. The judgement of grammaticality is incontestably subject to the level of tolerance for language uses that are "not flatly wrong". A demonstrative example for this is the use of the English word "ain't". The word "ain't" generally is considered to be a non-standard or informal contraction in English, and is generally not considered appropriate for formal writing. That being said, there may be certain contexts where the use of "ain't" could be considered acceptable or even appropriate, such as in certain types of creative writing, dialectical dialogue, or other forms of artistic expressions. Consequently, modern English dictionaries tend to include this entry but with an accompanying warning of usage. The above discussions indicate that in order to determine the grammaticality of a sentence, it is necessary to first specify the standards by which the sentence is being evaluated.

The judgement of grammaticality is also heavily dependent on the domain in which the language is used. For example, different sports have their own unique jargons that are considered as common phrases within their respective communities. A sentence like *He bunted the ball foul* would be considered perfectly acceptable and grammatical in the context of a baseball game, where "bunted" and "foul" are common terms that have specific meanings and are familiar to fans of the sport. In creative writing, writers may intentionally break conventional grammatical rules for artistic effect, and what may be considered as completely grammatically incorrect in standard English could be considered

acceptable or even desirable in this context. Thus, the judgement of what is considered "correct" or "incorrect" can vary widely depending on the linguistic context in which it is used.

As pointed out in Howcroft et al. (2020), the exact terms related to grammaticality vary throughout the literature, with alternative names such as "acceptability", "fluency" and "correctness" being present in the literature. To help clarify the meaning of these terms, I will provide a brief discussion of what they refer to.

In some NLG literature, the evaluation of "acceptability" is often used synonymously with that of "grammaticality" (Howcroft et al., 2020). However, "acceptability" and "grammaticality" are related but distinct concepts in linguistic theories. "Acceptablity" originates from the generative linguistic theory where the judgement of grammaticality is ultimately attested with the acceptability with its native speakers (Chomsky, 1957). This theory advocates that a language use has to be attested in the collective memory of its native users to be considered as being acceptable. An acceptable sentence with regards to a particular language is a sentence that is commonly acknowledged among a certain group of people who speak this language as their first language. The assessment of acceptability is also heavily influenced by the context in which language is used, including factors such as domain, linguistic registers and the level of formality in language use.

Whilst grammaticality emphasises the judgement of whether a sentence conforms to the rules of the grammar of a particular language, acceptability is concerned with how natural a sentence appears to a speaker of the language, regardless of its adherence to specific grammar rules. As such, acceptability is a gradient concept that characterises the degree to which a sentence aligns with the speaker's intuitive sense of what is typical or idiomatic in their language. In the influential syntax book, Chomsky (1957) presented the now well-known example of "colourless green ideas sleep furiously" (CGISF)[1]. This sentence posits an example that is instinctively grammatical but non-sensical to a native English speaker. The CGISF example was primarily used to highlight the distinction between syntax and semantics. Furthermore, the example serves to illustrate the notion that a grammatically correct language construction may not necessarily be acceptable or meaningful to a native speaker of the language. In turn, a sentence such as *Me and my friend went to the store* might be judged as grammatically incorrect, but it might still be acceptable in certain dialects or colloquial registers of English.

In line with this viewpoint, Lau et al. (2017) provided a detailed discussion of the relationship between grammaticality and acceptability. As per their assertions, grammaticality refers to "the theoretical competence that underlies the performance phenomenon of speaker acceptability judgements", whereas acceptability is what is measured "in experi-

[1]The CGISF was originally introduced in Chomsky's 1955 manuscript *The Logical Structure of Linguistic Theory*, which was published by Springer in 1975.

ments when we ask subjects to rate sentences". They acknowledge that grammaticality is fundamentally a syntactic concept, whilst acceptability can be affected by multiple factors, notably semantic plausibility, various types of processing constraints and others. This underscores the distinction between grammaticality and acceptability. Therefore, in this work, I employ the notion of **grammaticality** as a means of establishing a relatively objective reference for the quality of a text in terms of its adherence to certain standards of English grammar.

Fluency is another term that is frequently used in the literature. It is concerned with how smoothly and naturally a speaker can produce a language. In linguistic theories, *grammaticality* and *correctness* are sometimes used interchangeably to describe whether a sentence follows the rules of the grammar in a given language. However, this equivalence holds true only when the focus is solely on grammatical correctness, as in some linguistic contexts, the term "correctness" may have a broader connotation that encompasses not only grammaticality but also other factors of language use, such as semantic and pragmatic appropriateness.

3.2.2 Prescriptive vs descriptive

To evaluate the grammaticality of a text passage, it is essential to establish the standards against which it is to be evaluated. Existing methods for defining or describing grammaticality generally fall into two broad categories: *prescriptive* and *descriptive*. *Prescriptive* approaches typically hinge on pre-written rules that dedicate the preferred usage of language, as exemplified by dictionaries and grammar textbooks, which define word meanings and prescribe language construction rules. Conversely, *descriptive* approaches aim to observe and record how language is actually used. It seeks to provide an objective account of language use in reality, without bias to preconceived notions about how language ought to be used.

Similarly, the evaluation of grammaticality can also be approached from either a prescriptive or a descriptive standpoint. A prescriptive approach involves referencing a pre-existing set of rules that may be inflexible and outdated in some cases, lacking the ability to deal with nuanced cases. In contrast, a descriptive approach evaluates sentences by comparing them to representations of actual language use observed in real-life situations. This approach is more concerned with describing how language is used, rather than strictly adhering to a rigid set of rules.

The dichotomy between prescription and description continues to be a topic of ongoing debate in linguistic research (Cameron, 2005; Andrews, 2013). In the field of second-language acquisition, language prescription has long been criticised by modern linguists for its disproportionately normative practices and its reliance on predetermined establishments by linguistic registers and figures of authority (Swain and Canale, 1982; Pennington,

2002). Modern approaches to second-language teaching usually prioritise achieving fluency and emphasise communication in the language, rather than adhering strictly to a set of grammatical rules (Hinkel and Fotos, 2001). As a result, there has been a notable preference for descriptive methods in second-language education and publishing, as exemplified by the CEFR[2] and IELTS[3] assessment frameworks. Similarly, in the field of modern computational linguistics, a predominantly **descriptive** perspective has gained widespread acceptance. Common methods such as *generative grammars* and *data-driven language models* are both descriptive representations of language use. Generative grammars, such as phrase structure grammars, aim to systematically reflect language use in real-world contexts. Data-driven language models, on the other hand, are trained on large corpora to automatically extract language rules by screening and learning from real-world text. Both of these methods seek to model language use in natural contexts and provide effective representations of language phenomena, or as Santorini and Kroch (2007) phrased it, they are "intended as insightful generalisations" about actual language use. As a result, these two methods are both well-suited to serve as the foundation for grammaticality assessment. Specifically, grammaticality evaluation can be performed by assessing a sentence's adherence to either a thoroughly-designed formal grammar system or a pre-trained language model, which are referred to as grammar-based and data-driven approaches, respectively.

Both grammar-based and data-driven approaches have their own advantages and weaknesses. Choosing between these two approaches essentially comes down to a preference for either a comprehensive empirical coverage of all data, or a more profound explanatory understanding that may not be representative of all the possible data.

Grammar-based approaches measure grammaticality by comparing a sentence to a specific formal grammar, typically achieved through sentence parsing. This process involves analysing the sentence's structure and identifying how its components relate to each other according to the rules of the grammar system, allowing for a precise assessment of its grammaticality. Despite their precise nature, grammar-based approaches have faced criticism for relying on pre-existing formal grammars and parsers, as well as their perceived inflexibility in accommodating emerging language phenomena. However, it could be argued that their inability to fully account for novel or specialised uses of a language is not necessarily a drawback, particularly when considering their use as an evaluation tool. By adhering to established rules and structures, grammar-based approaches can provide a standardised framework for assessing the grammaticality of sentences and enabling systematic comparison across various linguistic phenomena. While they may not capture every aspect of language use, their adherence to formal grammar systems can help researchers establish a clearer understanding of the parsing process involved in sentence

[2]https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale.

[3]https://ielts.org/organisations/ielts-for-organisations/ielts-scoring-in-detail.

analysis. This level of granularity can be particularly useful in diagnosing subtle nuances and irregularities in language use, and in further refining existing formal grammars to better reflect real-world language phenomena.

In contrast to grammar-based approaches, data-driven approaches rely on training language models to learn the distribution of language from existing corpora. These models are designed to capture established patterns and precedents of language use, which can then be used to assess the grammaticality of previously unseen sentences. The main strength of data-driven approaches is that they reduce the need for manual labour in creating formal grammar systems for a particular language, and instead rely on large corpora to inform their models. Additionally, they allow greater flexibility in adapting to new use cases, provided that a sufficient amount of relevant data is accessible. However, this method is not without issues. Collected datasets often contain noise, which can complicate the identification of correct grammar usage. Conversely, most datasets tend to exclude certain instances of grammatical usage. This problem is exemplified by the previously discussed example of CGISF. Although the CGISF sentence is grammatically correct, it was likely not included in any known corpus when it was introduced. This raises concerns about relying entirely on statistical modelling of data from multiple collected corpora to make grammatical judgements, as grammatically correct sentences may not always be statistically probable. A further consideration is the use of manual annotations in many datasets. The annotation practice, as a whole, grants considerable influence to a restricted group of annotators who may have limited familiarity with linguistic theories. This issue is further compounded by the inherent diversity of annotators' underlying assessments of grammaticality, resulting in potential issues with data quality.

To enable the automation of grammaticality evaluation, it is necessary to establish a technical formal definition of grammaticality. In this thesis, I have opted for a **grammar-driven** approach to assess grammaticality, focusing on the acceptability (more specifically, parsability) with a generative grammar. There are two primary reasons for this choice. Firstly, for a syntax-based concept like *grammaticality*, it could be argued that a linguistics-inspired method would lead to a clearer and more comprehensible analysis of grammatical status. Secondly, relying solely on large amounts of data for grammaticality judgements may not be the most reliable approach, as the collected datasets often contain errors and inaccuracies, and the exposed usages may not always be entirely correct. Additionally, such models have been known to unfavourably encourage passive acceptance of misuses. Based on these considerations, I contend that for the particular evaluation of grammaticality, a concept with a relatively clear definition and a set of well-studied rules to refer to, adopting a grammar-based method is a sensible choice.

3.2.3 English Resource Grammar (ERG)

In the previous section, I discussed the distinctions between prescriptive and descriptive approaches regarding grammaticality judgement, and introduced two widely used varieties within the descriptive category: grammar-based approaches and data-driven approaches. In this thesis, I use the English Resource Grammar (Flickinger, 2000, ERG) for grammaticality evaluation. The ERG is an instance of the *grammar-based* approaches.

The ERG is a broad-coverage, high-precision grammar for English. Its development and maintenance encompass many variants of language structures. It belongs to the broader family of head-phrase structure grammar (Pollard and Sag, 1994, HPSG). The ERG was engineered to cover as many grammatical sentences as possible whilst correctly rejecting ungrammatical sentences. This is in contrast to statistical evaluation approaches which are trained on existing textual data, which will generally provide some analysis for a given text, regardless of its actual state of grammaticality.

The ERG is predominantly motivated by linguistic expert knowledge. However, the ERG's dependence on pre-defined generative rules does not limit its reflectiveness, as it is under active development to keep up with new vocabularies and language patterns. Despite this continuous maintenance, the ERG also offers stable releases to support research consistency and reproducibility. For the purposes of this research, all experiments were conducted using the 2020 release of the ERG. The ERG provides a vocabulary consisting of more than 30K unique lexical entries. The heuristic handling of unknown words in the ERG utilises part-of-speech (PoS) information and specifically designed regular expressions. Its curatable lexicon inventory and great tolerance to relaxation and mal-rules allow flexible adaption to any specific domains. The ERG has long been used to assess human-written texts. For instance, Flickinger (2011) tested the ERG on a variety of collections of English text and noted that the ERG had a coverage of 87.4% for English Wikipedia and 96.1% for E-commerce corpora. Similarly, Baldwin et al. (2013) used the ERG to examine the degree of syntactic noisiness in social media texts.

In this thesis, I employ the concept of **parsability** with the ERG as a *proxy* for **grammaticality** evaluation. In other words, a sentence is considered grammatically well-formed if a parse can be successfully obtained using the ERG. By contrast, a sentence rejected by the ERG or its relaxed versions is considered to be ungrammatical. The ERG is intended for applications where precision is required. By employing a well-defined, broad-coverage formal grammar to map grammaticality to parseability, it becomes possible to accurately specify the precise notion of "grammaticality" under investigation. This clarity is particularly valuable for diagnostic inspection, which is a primary objective of this study.

Parsing with the ERG usually yields one of the three cases:

(1) Parseable.

(2) Unparseable due to exhaustion of search space. The parser exhausted the entire search space of derivations for an input, and concludes that it does not have a derivation in the ERG.

(3) Potentially unparseable due to resource limitations. The parser reaches its limit of either memory or time without finding a parse for an input.

As previously mentioned, the outputs generated by the models evaluated in this thesis consist primarily of sentence-level units. The maximum length of these generated outputs is approximately 50 words. To prevent false negative failures caused by resource shortage (i.e., case (3) in the list above), I experimented with an Intel Xeon Gold CPU 6142 @2.60GHz and a RAM size of 100G, with a generous timeout threshold of 10 minutes for all experiments. These settings proved adequate for parsing all the sentences analysed in this thesis. Parsing the majority of sentences in this research took less than 1 minute and consumed less than 1GB of RAM. The maximum RAM consumption observed for all parsed texts was around 32GB. Consequently, it is reasonable to conclude that the parsing results in this thesis are restricted to cases (1) and (2), and that if a sentence does not obtain a parse with the ERG within the time threshold, it is considered unparseable by the ERG.

3.2.4 Existing work on automatic grammaticality evaluation

A substantial amount of research has been dedicated to examining issues related to text grammaticality. This section highlights the current body of work related to automatic grammaticality evaluation, with a specific emphasis on publications that *explicitly* evaluate grammaticality. It is important to note that metrics designed for assessing general text quality, where grammaticality is considered only one aspect among several others, will not be discussed in this section.

The task of grammaticality judgement is typically framed as a classification task (Cherry and Quirk, 2008; Post, 2011; Wagner et al., 2009; Warstadt et al., 2019; Cherniavskii et al., 2022), where classifiers are trained using datasets consisting of both naturally-occurring grammatical examples and negative examples that are ungrammatical. Negative examples are often artificially constructed using simple adversarial operations (such as insertion, deletion, or replacement) or through round-trip translation - that is, translating grammatical sentences into other languages and then back into the original language (Pauls and Klein, 2012; Bernardy et al., 2018). Early work in this area investigated feature-based approaches that utilised features derived from parser outputs and part-of-speech n-grams (Wagner et al., 2009). A similar idea was investigated in Mutton et al. (2007), where the outputs from multiple parsers, including the log probability of the most probable parse,

the number of invalid parses and the number of partially parsed trees, were combined to create discriminatory features for a feature-based grammaticality classifier.

Other works argue that the judgement of grammaticality should be viewed as a regression task, where gradient scores are assigned to indicate the degree of grammaticality achieved by a text. In light of this, Heilman et al. (2014) implemented a linear regression model for grammaticality scoring for L2 learner texts with a selection of linguistic features, in particular misspelt word counts, outputs from a variety of parsers and scores predicted by an n-gram language model. In their work, parsing results with the ERG are used as a minor feature in a statistical model, without any adaption to the domain. Along similar lines, Napoles et al. (2016) proposed a feature-based, reference-less metric to predict the gradient of sentence grammaticality for the task of grammatical error correction (GEC).

Other studies have suggested that evaluating grammaticality should be approached as a regression task, wherein the *likelihood* of a sentence with pre-trained language models is used as an indicator of its grammaticality. Such likelihoods are typically estimated using a variety of inherent language model parameters, with model perplexity being the most commonly used measure. Additional measures include negative cross-entropy and synthetic log-odds ratio (Pauls and Klein, 2012, SLR), the latter of which computes the log-probability of a sentence under a given language model, normalised by sentence length and word frequency. There are a wide array of language models that can be employed for this purpose, ranging from the traditional n-gram language models, Bayesian Hidden Markov Models (HMMs), and Latent Dirichlet Allocation (LDA) models (Lau et al., 2015; Niu and Penn, 2020), to contemporary neural network based language models such as the pre-trained BERT-based language model (Devlin et al., 2019; Lan et al., 2019; Zhang et al., 2019b).

An additional class of grammaticality evaluation methods is motivated by existing work on grammar error correction. These approaches are classified as edit distance based metrics as they gauge grammaticality by measuring the number of edits required to transform a generated sentence into a reference sentence. Common examples of these metrics include the word error rate and the Levenshtein Distance (Levenshtein et al., 1966). These evaluation approaches require previously collected reference data to compute the edit distances.

The most closely related work to the grammaticality evaluation in this thesis is the study conducted by Wei et al. (2018), which investigated the relationship between parseability with the ERG and grammaticality for output texts generated by sequence-to-sequence machine translation models (Sutskever et al., 2014). The authors asserted that the ERG's extensive coverage of linguistic phenomena renders its parseability an effective indicator of grammaticality in input texts. Similarly, I concur with their advocacy of the ERG as a tool for grammaticality evaluation. In fact, I began exploring this possibility around the

same time as their study. Whilst their work primarily focused on the feasibility of using the ERG for neural machine translation, this thesis incorporates the ERG component into a broader framework of GFD evaluation.

3.3 Faithfulness

The primary objective of natural language generation (NLG) is to generate textual output that effectively conveys a specific message or information. Thus, it is of utmost importance that the generated text faithfully represents the content in the input text. Faithfulness serves as a metric to measure the extent to which a text generated by a model reflects the intended meaning of the input data. Faithfulness is crucial for real-world applications that rely on NLG systems. For example, in the fields of medicine and law, inaccuracies in the generated text can have significant ethical and legal implications, potentially leading to serious consequences.

Evaluating the faithfulness of generated text is therefore critical to guarantee that the text accurately communicates the intended purpose of the generation task and is free from errors or misunderstandings. Furthermore, by improving metrics and methods used to evaluate faithfulness, researchers can gain a better understanding of the strengths and weaknesses of different generation techniques, and develop new approaches that produce more faithful texts.

In this section, my aim is to provide a clear explanation of the concept of faithfulness that I will be evaluating in this thesis. I will also explore similar concepts that have been used interchangeably with faithfulness in existing literature. Additionally, I will discuss the concept of hallucination, which is closely associated with faithfulness in the context of language generation. Finally, I will provide an overview of the current research on automatic faithfulness evaluation.

3.3.1 What is faithfulness?

The discussion around the concept of faithfulness in linguistics has its roots dating back to at least the 1700s. Herder, in his early exploration of translation theory (Herder, 1767), emphasised the importance of semantic faithfulness as the primary goal of translation. To achieve this, he advocated for an "accommodating" approach, where the content of a translation is made to closely conform to that of the source text.

In modern NLG research, faithfulness has been extensively examined in tasks such as machine translation (Weng et al., 2020; Feng et al., 2020) and text summarisation (Cao and Wang, 2021; Chen et al., 2021; Aralikatte et al., 2021). In these tasks, the generation of model outputs is strictly grounded by the input content, making faithfulness an essential criterion for evaluation. The faithfulness evaluation proposed in this thesis is intended for

general NLG scenarios. While the specific implementation of the faithfulness evaluation may be tailored to adapt to different task scenarios, evaluation in this work is uniformly focused on two key aspects - the generated text should be faithful to (1) the underlying requirements of the generation task itself, and (2) the content of the input source. In this thesis, I refer to the first faithfulness aspect as *task faithfulness*, and the second aspect as *grounding faithfulness*.

The requirement for evaluating task faithfulness in NLG is to ensure that the generated text is aligned with the inherent requirements of the task. This stipulation is relatively straightforward to comprehend. For example, if the NLG task is to generate a summary for a news article, the generated text should contain the key points and important details of the article, and present them in a concise manner. If the NLG task is to generate a weather report, the generated text should effectively convey the current weather conditions and the forecast for the coming days. In other words, the generated text should meet the specific requirements of the task at hand. Given that contemporary NLG research has primarily focused on developing specific model architectures to meet particular task specifications, such as the Moses system for machine translation (Koehn et al., 2007) and DialoGPT for dialogue generation (Zhang et al., 2020b), it is reasonable to assume that most existing models meet this requirement.

The requirement of grounding faithfulness emphasises that the generated text should remain faithful to the input content upon which the generation process is grounded. Arguably, every generation task is inherently constrained by some form of grounding in the input. For example, in machine translation, the information contained in the source sentence must be effectively conveyed during the translation process. Similarly, in image captioning, the input image serves as the grounding for the generation process. While having direct access to the ground truth is ideal for conducting faithfulness evaluations, most NLG tasks lack complete or direct access to ground truth. In such cases, it is necessary to identify underlying representations that suitably reflect the information contained in the input. Such representations are commonly referred to as *world models*. Once the world model has been established for a particular generation task, it can be used to assess the degree of faithfulness in the model outputs. In faithfulness evaluation, the input information and the output information are both converted to symbolic representations, which are compared to decide if they are equivalent. If so, the output can be considered as faithful to the input content.

3.3.2 Faithfulness and related concepts

In existing literature, the concept of faithfulness is often used synonymously with related terms, in particular "factuality", "appropriateness", "correctness", and "goodness". I will provide an overview of the connections and distinctions between these concepts.

Factuality is a concept that is frequently used in NLG literature (Gabriel et al., 2021; Goyal and Durrett, 2020), measuring whether the content of a text is factually true or not. In other words, it refers to the degree to which the generated text reflects reality or conforms to known facts. I argue that the concept of *faithfulness* differs from *factuality* in that faithfulness is grounded in the notion that outputs should be firmly based on the underlying purpose of the generation and the contextual facts which are presented in the input. In contrast, *factuality* implicitly contains a requirement for consistency between an output and general world knowledge, the evaluation of which inevitably requires complicated constructions of knowledge bases to represent real-world common sense, and is therefore beyond the scope of this thesis. Therefore, while factuality is an important consideration in NLG, it is not the primary focus of this thesis.

Appropriateness refers to the degree to which an output is appropriate in the given contexts (Webb et al., 2010; Horbach et al., 2020). The judgement of appropriateness is inherently subject to factors such as the register and domain of interest. *Goodness*, as defined in Howcroft et al. (2020), refers to the degree to which an output is good. Although these two aspects share similarities with the concept of faithfulness, objectively assessing whether a model's generation is appropriate or good can be challenging because these terms are highly generic and subjective in nature.

The term *correctness* has also been used to measure properties similar to faithfulness (Li et al., 2018; Zhang et al., 2020c). Correctness measures the degree to which outputs are correct. Evaluations of this criterion typically ask whether an output is correct or not given the context. However, correctness includes not only content consistency, which is similar to what faithfulness measures, but also form-level correctness and several other factors. Therefore, it is essential to provide further specification when using this term.

Based on the above considerations, I have chosen to use the term "faithfulness" to describe the particular aspect of NLG evaluation that I am proposing in this thesis. While there are other terms that have been used in the literature (such as "factuality"), I believe that "faithfulness" better captures the specific focus of my evaluation framework. This reinforces my previous argument that the selection of terminology for NLG evaluation should be conducted with great caution to avoid potential confusion and to allow empirical comparisons across different papers.

3.3.3 A brief note on hallucination

Deep neural networks are known to be prone to the undesirable phenomenon of *hallucination*, as has been observed in many existing works (Lee et al., 2019; Rohrbach et al., 2018). Hallucination in psychology is defined as a perceptual experience where someone has a visual experience as of an object, but is actually not seeing an object at all (Bleuler and Brill, 1924; Hinsie and Campbell, 1970). Although there is not a commonly agreed

formal definition of hallucination in the context of NLG, the term is typically used to refer to the phenomenon in which models generate seemingly persuasive information which is unfaithful to the provided source content (Ji et al., 2022). Hallucination is closely related to the concept of faithfulness, and can be considered as an opposite to faithfulness. In other words, a model output that is confident but unfaithful to the source content can be considered a hallucination.

It is crucial to emphasise the distinction between a hallucination and a plausible inference, as they represent two distinct aspects of language understanding. A hallucination refers to a situation where a model generates or produces content that is not supported by the input or context, leading to unreliable or misleading information. On the other hand, a plausible inference signifies a valuable reasoning ability that we expect models to possess. Plausible inference involves determining whether a natural language statement logically follows the given information. This reasoning ability has received significant attention in text entailment research, where researchers have studied methods and techniques to effectively capture the logical entailment, contradiction, or neutral relationship between pairs of statements (Dagan et al., 2005; Bos and Markert, 2005; Bjerva et al., 2014; Bowman et al., 2015; Poliak, 2020; Alharahseheh et al., 2022).

Many instances of hallucination have been observed in the outputs of various NLG models, including those designed for machine translation (Koehn and Knowles, 2017; Lee et al., 2019; Raunak et al., 2021), data-to-text generation (Parikh et al., 2020; Wiseman et al., 2017; Dhingra et al., 2019; Nie et al., 2019; Wang, 2019) and summarization (Huang et al., 2021; Maynez et al., 2020). The frequent occurrence of hallucination in neural NLG highlights the necessity of incorporating a faithfulness evaluation into the evaluation checklist, as well as for improved automatic evaluation metrics, especially when the concerned approaches are neural networks.

Existing studies have explored various methods to reduce data noise and mitigate the problem of hallucination. These methods include augmenting an existing corpus with additional information, and filtering out data pairs that are considered mismatched. Data augmentation techniques (Fan et al., 2019; Gunel et al., 2020; Huang et al., 2020; Chen et al., 2021; Bi et al., 2019) involve adding new training examples or incorporating additional contextual information to existing data, which can help reduce the chances of generating hallucinated text. Alternatively, corpus filtering techniques (Raunak et al., 2021; Liu et al., 2021b; Shen et al., 2021) involve removing examples that contain errors or inconsistencies, which can improve model performance by reducing the likelihood of generating incorrect or irrelevant outputs. Additionally, researchers have also developed new datasets with hallucination reduction in mind. One such example is the TOTTO dataset (Parikh et al., 2020), which was proposed specifically for table-to-text generation. Another line of exploration addresses the issue of hallucination by modifying the current

structure of generation models, either by incorporating an additional encoder that explicitly models the fact tuples in the input source (Cao et al., 2018; Huang et al., 2020), or by applying purposefully designed attention mechanisms that force the model to condition the generation process on the relevant knowledge contained in the source documents (Wu et al., 2021; Aralikatte et al., 2021). These approaches have shown promise in reducing hallucination and improving the overall quality of generated text.

3.3.4 Existing work on automatic faithfulness evaluation

In light of the varied use of terminology related to faithfulness (as discussed in section 3.3.2), I will discuss existing research that explores faithfulness and its associated concepts, in particular *factuality*, *factual consistency* and *factual accuracy*.

The evaluation of faithfulness in text summarisation has been extensively explored in existing research. One common approach involves using question answering (QA) models. This QA-based method generates questions based on the source documents and then uses QA models to generate answers based on the summary (without access to the source text). The faithfulness of the summary is assessed by examining whether the generated questions can be correctly answered using only the information in the summary. This approach transforms the faithfulness evaluation of a model-generated summary into the evaluation of the extent to which the summary provides sufficient information to answer questions proposed based on the source document. Some recent studies, such as Honovich et al. (2021) and Scialom et al. (2021), have applied this QA-based approach to knowledge-grounded dialogue and text summarisation tasks. Furthermore, Honovich et al. (2022) proposed a meta-evaluation protocol for this method which covers various tasks, including text summarisation, knowledge-grounded dialogue, fact verification and paraphrasing. However, this approach requires a separate question generation component and a question answering component, and the creation of new datasets with manual annotations is often necessary to enable this type of evaluation.

Another research approach involves using textual entailment based methods. For instance, Goyal and Durrett (2020) employed dependency arc entailment (DAE) to assess the faithfulness of summaries and paraphrases in tasks of summarisation and paraphrasing. In this method, a dependency arc in the dependency tree of a generated text is deemed entailed by the source if the semantic relationship between its head and child (such as "nsubj", "nmod:about", "amod") is entailed by the source sentence. In this way, dependency arcs are treated as semantic units that can be interpreted independently. This approach decomposes the evaluation of faithfulness of model-generated text into the evaluation of individual components of their structured representations (in this case, dependency trees).

Various studies have explored entity-centric methods for evaluating the faithfulness of generated text, which involve extracting information entities and subsequently verifying

their authenticity in relation to the source information (Novikova et al., 2017; Rohrbach et al., 2018; Liu et al., 2021b). One recent metric proposed by Goodrich et al. (2019) employs a two-step pipeline to extract fact tuples consisting of entities and their accompanying relationships. The pipeline first extracts all named entities in the sentence to be evaluated, and then classifies the relations between the extracted entity pairs. The authors also experimented with an end-to-end transformer-based model for fact tuple extraction. The resulting faithfulness evaluation is performed by measuring the precision of the set of relation tuples inferred from the generated text with respect to those of the ground truth text. Although the faithfulness evaluation approach used in the FHiG task in section 6.5 follows a similar vein, it differs from the techniques suggested in other entity-centric studies, which are specific to their respective application domains and rely on assumptions about the nature of the input content and the structure of the sentences used to express that content. While the high-level perspective of these techniques is useful for that particular task, I advocate for generalising the faithfulness evaluation and incorporating it into a larger evaluation framework that can be applied to a wide range of tasks, provided that an underlying world model of the ground truth is available.

In recent years, fact-checking has become a significant research area, with a growing number of studies dedicated to benchmarking systems and evaluating their effectiveness. For example, the FEVER benchmark (Thorne et al., 2018, 2019) and similar tasks have garnered significant attention in this field. A number of studies have explored meta-evaluation for metrics related to faithfulness. Meta-evaluation involves assessing existing evaluation methods to ensure their accuracy and reliability. For example, Gabriel et al. (2021) introduced a meta-evaluation framework for various factuality metrics proposed for summarisation tasks.

3.4 Diversity

While grammaticality and faithfulness are crucial requirements for language generation, focusing solely on these elements can lead to exploitation by using generic phrases that are often true. For instance, in the case of image captioning, descriptions such as *There is an object* or *A minimum of zero people are in the picture* (which is technically correct even without any individuals present) are grammatical and faithful captions, but exhibit excessively generic comments that could be applicable to nearly all images. In addition, recent studies indicate that while maximum likelihood estimation (MLE) training prompts generation models to create texts with a high likelihood within the training set, these models frequently produce monotonous and generic texts due to this method (Dai et al., 2017; Jiang and de Rijke, 2018; van Miltenburg et al., 2018). These insights motivate the third fundamental requirement of NLG output to be diverse.

This section discusses the concept of *diversity* in text, which can be broadly categorised into three types: lexical diversity, syntactic diversity, and semantic diversity. I will first discuss the three dimensions of diversity, followed by explanations of the two diversity measures that are implemented in this thesis - word-level diversity and parse-level diversity. In addition, I will provide a synopsis of prior work that analysed the three dimensions of diversity.

3.4.1 The multi-faceted nature of diversity

Diversity, particularly within textual content, can be examined through various lenses. Broadly, this examination can be divided into three categories: lexical diversity, syntactic diversity, and semantic diversity. Each of these categories provides a unique perspective, collectively contributing to a multidimensional understanding of text diversity.

Lexical diversity essentially pertains to the variance in word choice within a text. Texts that exhibit high lexical diversity use a wide range of vocabulary, thereby enhancing the richness and nuance of the passage.

Syntactic diversity shifts the focus to the diversity in syntactic structures within a text. Techniques like parsing and part-of-speech tagging are typically employed to analyse syntactic diversity. A high level of syntactic diversity tends to result in more dynamic and engaging text, as it prevents the overuse of certain sentence structures.

Semantic diversity, on the other hand, relates to the variety of meanings or topics covered in a text. A text with high semantic diversity often covers a wide array of topics or ideas, leading to more engaging and insightful text. However, assessing the semantic dimension of diversity is generally deemed challenging due to the inherent difficulty in evaluating the similarity and divergence of semantic meanings across multiple texts.

In this thesis, the evaluation of diversity within the GFD framework is primarily confined to lexical and syntactic diversity. More specifically, a *word-level diversity* is examined as an instance of lexical diversity, and a *parse-level diversity* as an instance of syntactic diversity.

The word-level diversity examines the variation in word usage within a text, defined as the normalised type-token ratio (Richards, 1987; Covington and McFall, 2010) of the text. Whilst word-level diversity examines lexical difference, parse-level diversity focuses on structural variety within model outputs. This facet of diversity specifically analyses the various language constructions that a model is able to generate, utilising derivation trees (Ivanova et al., 2012) obtained with ERG parsing (Flickinger, 2000).

Both the word-level and the parse-level diversity evaluations are versatile and applicable across various generation scenarios. Importantly, these two diversity measurements are both reference-less, circumventing the need for manually curated reference sets. The diversity evaluations within the GFD framework are particularly advantageous when

references are unavailable or the underlying construction rules are not readily accessible during evaluation. Details regarding the implementation of the two facets of diversity measurement for specific tasks are elaborated in chapters 5-7.

3.4.2 Existing work on automatic diversity evaluation

Although there is a lack of principled method for evaluating the diversity of NLG systems, the concept of diversity has been assessed in various ways in the existing literature, primarily classified into three dimensions: lexical diversity, syntactic diversity, and semantic diversity.

Lexical diversity. The most fundamental measure of lexical diversity is the type-token ratio (Richards, 1987, TTR), which quantifies diversity by calculating the ratio of unique words (types) to the total number of words (tokens). Various other measurements, such as the measure of textual and lexical diversity (McCarthy, 2005, MTLD) and moving-average type-token ratio (Covington and McFall, 2010, MATTR), provide methods to assess the range and variety of vocabulary utilised in a text.

Syntactic diversity. A multitude of evaluation methods have been proposed in recent years to assess syntactic diversity. Li et al. (2016) introduced the DIST measure, defined as the ratio of distinct n-grams in evaluated sentences. The usual considerations for this method include unigrams, bigrams, and trigrams, and it has been widely adopted in numerous studies (Nakamura et al., 2018; Zhang et al., 2018b; Xu et al., 2018; Gao et al., 2019a). Zhang et al. (2018b) further proposed an Entropy metric, a modification of the DIST measure, which accounted for the frequency difference of n-grams.

In another approach, Zhao et al. (2017) designed a syntactic diversity measure using an inverse version of BLEU, where the generated texts are taken as the "reference" for the BLEU calculation, and the actual references provided in the test set are taken as the "hypothesis". They argued that a higher score in this context implies greater diversity in the generated texts. Similar methodologies were also employed in subsequent studies (Shi et al., 2018; Gao et al., 2019b).

Semantic diversity. Several studies have recently ventured into developing evaluation methods for semantic diversity. Wang and Chan (2019) introduced an automatic metric to measure semantic diversity in image captioning, based on latent semantic analysis (LSA) which was kernelised using CIDEr similarity. A high evaluation score in this measure indicates a varied set of topics in the evaluated texts, and therefore a high semantic diversity. Other approaches include the NLI Diversity (Stasaski and Hearst, 2022), which leveraged pretrained natural language inference (NLI) models, and the Sem-Ent metric (Han et al., 2022), which measured the semantic diversity of generated responses by mapping these responses into a semantic latent space using pre-trained language models.

Despite the recognised importance of semantic diversity in model evaluation, it is important to re-emphasise that the proposed GFD framework predominantly focuses on

the lexical and syntactic facets of diversity. The primary objective of the GFD framework is to serve as an initial effort towards a clearly-defined, diagnostic tool for NLG evaluation. The selection of the metrics is thus steered by the practicality of finding an appropriate method to operationalise them. As such, lexical and syntactic diversity was emphasised in the GFD framework due to this practicality consideration.

Chapter 4

Rationale for the choice of NLG tasks

The main objective of this thesis is to propose a structured, diagnostic evaluation framework for natural language generation tasks. To apply the GFD evaluation framework proposed in section 3, I have carefully chosen three different NLG tasks for examination: image captioning, data-to-text generation, and dialogue generation. Within the general framework of these three tasks, I specifically explore three variants of each generation scenario, namely synthetic image captioning, football highlight generation, and topic-shift dialogue generation.

The objective of this section is to present the three NLG tasks and the particular variants that are being investigated in this thesis, and provide the rationale behind the selection of these tasks, and especially the variants. I will elaborate on my decision to select image captioning as the starting point for GFD evaluation, as well as why I specifically opted for a synthetic setup to investigate the evaluation of this task. I will discuss the reasoning behind selecting football highlight generation as a particular case of data-to-text generation, as well as the decision to explore topic-shift dialogue generation for dialogue systems. I will then introduce how the other two tasks, football highlight generation and topic-shift dialogue generation, present natural next-level challenges for exploration following the image captioning task.

4.1 A brief summary of three NLG tasks

This thesis investigates three different NLG scenarios: image captioning, data-to-text generation, and dialogue generation. In section 2.2.1, I provided a brief introduction to these tasks as well as other NLP tasks. To facilitate discussion, I will provide a concise summary of the three tasks below.

- *Image captioning* involves generating textual descriptions of images. In this task, the input is an image and the goal is to generate a natural language description of

the image. Image captioning is commonly used in image search engines and social media platforms where users can search for images using keywords.

- *Data-to-text generation* involves generating natural language text from structured data. In this task, the input is structured data, such as a table, diagram or database, and the goal is to generate a natural language text that summarises the data or presents it in a more readable format. Data-to-text generation is commonly used in financial reports, weather forecasts, and sports summaries.

- *Dialogue generation* involves generating natural language responses in a conversational setting. In this task, the input is a conversation between two or more participants and the goal is to generate a natural language response that continues the conversation. Dialogue systems are used in chatbots, virtual assistants, and customer service applications.

These three tasks are widely studied in the field of NLG. All three tasks come under the broad category of natural language generation, which involves producing natural language text from various types of input data. This shared characteristic leads to similarities such as the requirement to produce coherent and grammatically correct text. However, they differ significantly in terms of their intrinsic task requirements and focuses, thus providing a good spread of NLG scenarios.

4.2 Three particular variants

It should be noted that for each task, I have selected a specific scenario to research on within the larger context of these three tasks. More specifically, for image captioning, I examined the synthetic scenario composed of abstract visual scenes made up of coloured shapes. For data-to-text generation, I chose the English Premier League football as the domain of interest and focused on generating match highlights from match statistics. For dialogue generation, I focused on the particular phenomenon of topic shifting and aimed to model the behaviour of topic shifts in dialogue generation.

In this section, I will provide detailed explanations of the specific cases chosen for each NLG task, taking into account the particular constraints posed by each case scenario[1]. Specifically, I will explain why I chose the synthetic scenario for the image captioning task and how it facilitates the core aim of this thesis. I will then discuss the details of football highlight generation and how it relates to the general definition of data-to-text generation, as well as how the additional constraints impact the generation task. Finally,

[1]Note that the formal definitions of each task will be introduced in their respective chapters, namely chapter 5 for synthetic image captioning, chapter 6 for football highlight generation, and chapter 7 for topic-shift dialogue generation.

I will introduce a specific case of dialogue generation, where the focus is on modelling the phenomenon of topic shifting in utterance generation during dialogue turns. When the three NLG *tasks* are mentioned in the rest of this thesis, reference is made to the three *variants* with specifically defined task constraints, unless otherwise stated. Furthermore, it is critical to emphasise that all three tasks focus on short-form generation. The target generation candidates typically consist of only a few sentences, with the majority of them containing only one sentence.

Synthetic image captioning. In this thesis, I examined image captioning under a synthetic scenario, where visual scenes are made up of abstract coloured shapes. Objects in the image were randomly sampled from predefined shape and colour options, and the combination of multiple objects and their relative positions were also randomly selected. This synthetic setup provided a precise control over the generated visual scenes, which is particularly helpful for the GFD evaluation. Additionally, the random rendering of image content allowed for a diverse inventory of possible visual images, facilitating the study of image captioning under different visual constructions.

Football highlight generation. Following the synthetic image captioning task, I explored data-to-text generation, specifically in the domain of football highlight generation. The objective of this task is to generate a concise highlight of an event during an English Premier League football match, based on the structured data records of the *match results*, which I interchangeably refer to as *match statistics*. This task represents a natural progression from the synthetic image captioning task, as it involves real-world input information that is structured but not entirely controlled. This allows for a level of generalisation while maintaining a reasonable degree of control for diagnostic evaluation of a real-world task.

Topic-shift dialogue generation. The previous two tasks involved generating textual outputs from structured input contents. Unlike these two tasks, the dialogue task is a text-only generation task. To expand the scope to an open-domain free-form generation scenario, my investigation proceeded to explore topic-shift dialogue generation. This task represents a specialised form of dialogue generation, with emphasis on modelling the topic shifting behaviour commonly observed in human conversations. The open-domain chit-chat nature of this task ensures that the generation occurs in a considerably broader candidate space. The quality of a generated utterance is judged by how well the utterance responds to the previous conversation history. There can be a large number of possible realisations spanning diverse topics and fields. Consequently, the generation of the next utterance is not as strictly constrained by the input content compared to the other two tasks. However, by imposing a topic-shift constraint, the generation process is explicitly required to produce topic-shift utterances in response to the dialogue history. This imposes a strong grounding in the task target itself, thus forming an intriguing combination of grounding to inspect and a much more challenging setting for the GFD evaluation.

These tasks are diverse in nature and involve a wide range of language generation scenarios, which present unique challenges for evaluating their performance. Therefore, it is important to carefully consider the specific characteristics of each task with regards to *level of grounding from input*, *task constraint* and *language use*, and how these aspects might affect the GFD evaluation process.

Level of grounding from input. The process of generating textual output is typically dependent on the content contained in the input information. The three tasks provide varying levels of grounding forms. For instance, in image captioning and data-to-text football highlight generation, the generated text must strictly align with the content of the input information, whether it is represented in the form of images or match statistics. This necessitates a high level of grounding to ensure that the generated text accurately corresponds to the input information. In contrast, in open-domain chit-chat, the generated output is expected to be coherent with the overall conversation context. While this level of grounding is still important, it is more relaxed compared to the other two tasks, as the generated response must simply maintain relevance to the conversation history rather than strictly aligning with the input content.

The degree of grounding required for each task has a significant impact on the design of evaluation strategies. The strict grounding required in image captioning and data-to-text generation tasks necessitates more rigid evaluation strategies to ensure that the generated text aligns with the input information. In contrast, the more relaxed grounding constraints in open-domain chit-chat may require more flexible evaluation strategies to evaluate the relevance of the generated text in the context of the conversation history. In such cases, evaluation metrics may need to focus on the coherence and relevance of the generated text, rather than on its alignment with the input information.

Task constraint. As demonstrated by the three tasks evaluated in this thesis, there are notable differences in their respective task purposes. For instance, image captioning involves generating a textual description that accurately corresponds to the visual content of an image, while football highlight generation requires the generation of a concise summary from structured data records. Topic-shift dialogue generation involves generating responses that naturally change the topic of an ongoing conversation given a dialogue history. Furthermore, the distinct requirements of each task introduce varying sizes of potential output spaces. Generating topic-shift responses results in a relatively large potential output space, as there are various possible directions to take the conversational topic. The generated text can vary widely, making it challenging to evaluate the quality of the generated text accurately. In contrast, the potential output space for the data-to-text generation task is relatively small, as the output text must closely adhere to the specific data records. Therefore, it is crucial to account for the unique task constraints and their corresponding potential output spaces when applying the GFD evaluation framework to

each task.

The aforementioned two factors have a significant impact on the evaluation of **faithfulness**, as the level of grounding in the generated text can directly affect the implementation of faithfulness evaluation. Faithfulness essentially examines two aspects: (1) the degree to which the generated output is faithful to the input information, and (2) the degree to which the generated output adheres to the requirements of the task itself (i.e., whether the requirements are met or not). In the case of topic-shift dialogue generation, faithfulness may require the generated response to be grounded in the previous conversation history whilst also naturally shifting the conversational topic. Therefore, the requirements for faithfulness can vary significantly between tasks, and it is crucial to design evaluation strategies that account for these specific factors to accurately assess the quality of generated text.

Language use. The three language generation tasks examined in this study are diverse in nature and belong to different domains, which results in the use of distinct domain-specific vocabulary and language styles. For example, the first task requires the generation of text that describes simple visual scenes using a controlled vocabulary, whereas the second task generates concise texts from real-world semi-structured data in a specific football domain. In contrast, the third task generates text in an open-domain chit-chat setting, but with a specific topic focus specified by the task. This variation in domain-related language use has a crucial impact on the evaluation of **grammaticality**, as the evaluation of grammaticality requires the consideration of linguistic norms and patterns specific to the domain for an accurate assessment. This highlights the importance of domain adaptation when evaluating the quality of language generation systems. The evaluation process should carefully take into account the specific context in which the generated text is intended to be used to ensure that the evaluation metrics accurately reflect the quality of the generated text.

In summary, we can observe a common thread that ties the three tasks together when zooming out from specific task features. This thread is the level of grounding of the input information imposed on the generation process, which ranges from the highly grounded description-focused captioning and highlight generation tasks to the more loosely-formed dialogue generation task. Additionally, the availability of the underlying ground truth decreases from the fully controlled synthetic image captioning setup, to the partially accessible match statistics and finally the implicitly referred dialogue context.

Examining each task individually allows us to discern how task-specific differences, such as the number of input modalities, introduce new dimensions to the generation and evaluation processes. While it may be challenging to establish a standard practice for evaluating natural language generation models, adopting better-designed, finer-grained evaluation practices holds potential for improving the field. This ultimately will lead to

a more focused evaluation process, and enable more meaningful comparisons between different NLG models.

Chapter 5

ShapeWorldICE: Synthetic image captioning and evaluation

5.1 Introduction

The first task that I investigate is synthetic image captioning[1]. To begin, I will provide an introduction to this task and existing work related to the task (see section 5.2), followed by an introduction to the ShapeWorldICE benchmark created for this work (see section 5.3), and an exploration of the GFD framework proposed to evaluate this task, the evaluation results and detailed discussions (see sections 5.4-5.6).

5.2 The task of syntactic image captioning

Image captioning is a multimodal task that has drawn much interest in recent years. In the general form of image captioning, a model is presented with an image and the goal is to generate a natural language description for the image. More formally, given an image x, the image captioning task aims to generate a natural language sequence $y = y_1, y_2, \ldots, y_N$

[1]Acknowledgement: The results presented in this chapter are a continuation of the work the I did in the first year of my PhD. The content of this work was accepted and published as a conference paper with the title "Going Beneath the Surface: Evaluating Image Captioning for Grammaticality, Truthfulness and Diversity" at the MetaEval workshop of the AAAI 2020, co-authored with Thomas Sherborne, Alexander Kuhnle and Ann Copestake (Xie et al., 2020).
Preliminary work on using the ShapeWorld framework for captioning was carried out by Thomas Sherborne in an MPhil project co-supervised by Alexander Kuhnle and Ann Copestake. Sherborne's main experiments were on the very simplest ShapeWorld datasets (mainly on OneShape) since his evaluation methodology could not be extended. I co-developed the GFD evaluation for the image captioning task with Alexander Kuhnle. Kuhnle developed the ShapeWorld data generation framework, which served as the foundation for the diagnostic ShapeWorldICE dataset presented in this chapter. He also proposed the ShapeWorld-specific diversity evaluation approach for the task. Nevertheless, I extended the diversity evaluation by incorporating two measures, word-level diversity and parse-level diversity, which offer better generalisation to other generation scenarios beyond ShapeWorld. All other experiments and findings presented here represent my own work.

A few boats that are in a small body of water.
Many people are in boats sailing down a river.
A boat travelling along a river surrounded by grass fields.
A view of a river with a few boats on it.
People on a barge paddling down the river.

Two wooden benches sitting on a field of grass.
Cat sleeping on wooden park bench on grass near stone wall.
A cat sits on a wooden bench in the grass.
A wooden bench is sitting in a grassy area.
A park bench that has a cat on it.

Figure 5.1: Example images and reference captions in the MS-COCO dataset.

Caption 1: A *shape* is to the left of a semicircle.
Caption 2: A *blue triangle* is to the left of a semicircle.
Caption 3: A semicircle is below a gray triangle.
Caption 4: A semicircle is to the left of a triangle.

Figure 5.2: ShapeWorld example: spatial statements in the context of multiple shapes. The first three statements are faithful and diverse descriptions of the image. The fourth statement is wrong, but nonetheless exhibits a high degree of n-gram overlap with the faithful reference captions.

which describes the visual contents in image x. Figure 5.1 shows two example images and human-annotated captions in the MS-COCO dataset (Lin et al., 2014). Note that in this section, the terms *caption* and *description* are used interchangeably.

This task is not only significant but also technically challenging, as it requires a precise grasp of both the content of and the correlation between vision and language. Driven by recent advances in deep learning techniques along with the release of large-scale datasets specifically curated for the image captioning task, many recent image captioning models have been reported as achieving "super-human" performance on this task (Zhou et al., 2020; Hu et al., 2022; Demirel and Cinbis, 2022). However, researchers have observed that models tend to exploit the unfavourable biases present in the training data to game the generation task and achieve superficially impressive evaluation results, whilst their fundamental ability to understand images and generate text remains questionable (Mohamed et al., 2022; Bhargava and Forsyth, 2019; Zhao et al., 2021; Bakr et al., 2023). This is unsurprising for two main reasons. Firstly, the majority of existing real-world image captioning datasets

are collected through crowd-sourcing, which almost always compromises the quality of the collected data due to both the inherent limitations of the collection approach itself and the varying levels of expertise among the crowd annotators (Aker et al., 2012; Hossain and Kauranen, 2015; Daniel et al., 2018). Secondly, current evaluation methods for image captioning primarily rely on automated evaluation metrics such as BLEU, supplemented by coarse-grained human evaluations. Overall, there is an abundance of unfavourable biases hidden in many existing image captioning datasets, and a general inadequacy of current evaluation benchmarks to capture the crucial aspects of a model's performance on a specific task. As such, it becomes necessary to conscientiously explore alternative automated standards and accompanying datasets that can genuinely assess the actual performance of models.

To address the above two challenges in image captioning evaluation, I constructed a synthetic dataset specifically tailored for image captioning, utilising the publicly available ShapeWorld data generation framework (Kuhnle and Copestake, 2017). By focusing on a **synthetic** setting, it is possible to obtain a more comprehensive understanding of image captioning models. This synthetic dataset brings several significant advantages in terms of model evaluation:

1. Control over bias. By constructing a synthetic dataset, it is possible to carefully control unintended biases presented in the data. This enables a more balanced and controlled environment to evaluate image captioning models, mitigating the influence of unfavourable biases that may be prevalent in real-world datasets.

2. Availability of ground truth representations. In a synthetic setting, we have the advantage of knowing the ground truth representation behind every image. This enables a **reference-less** evaluation of model performance, as we can directly assess model-generated captions against the underlying ground truth representation of the images, rather than a couple of human-written reference captions.

3. Systematic evaluation. With a synthetic dataset, we can systematically vary the properties of the images and captions, which allows us to investigate the robustness and adaptability of image captioning models across different scenarios.

4. Generalisation ability. Using a synthetic setting allows us to design specific scenarios to test the generalisation ability of particular models. By evaluating model performance on unseen images generated under controlled conditions, we can gain deeper insights into how well the models can handle novel inputs.

A more comprehensive discussion of the rationale behind using synthetic benchmarks for fine-grained evaluation can be found in Kuhnle (2020, pp. 35-36). While I acknowledge the limitations of employing an entirely synthetic specification, such as the inherent

dissimilarity between synthetic and real-world data, I believe that the synthetic nature of the data does not diminish its utility for my purposes. The primary objective of this synthetic task is to facilitate a more comprehensive evaluation of image captioning models. The synthetic dataset coupled with the GFD evaluation benchmark serves as a supplementary approach to mainstream evaluation practices. As such, it should not be considered as the perfect test-bed for assessing model performance, but rather as an essential preparatory test that models must pass to demonstrate a certain level of genuine image understanding and linguistic composition.

5.2.1 Related work on image captioning

5.2.1.1 Existing datasets

A wide range of *real-world* datasets have been used for benchmarking image captioning models and evaluating their performance. These datasets usually contain complex scenes from the real world, with each image associated with multiple captions in which detailed information about the image is described. An early attempt at collecting image captioning data is the Flickr8k dataset (Hodosh et al., 2013), which consists of 8k images collected from Flickr. Each image is paired with five different crowd-sourced annotations that describe the entities and events depicted in the image. Flickr30k (Young et al., 2014) extends the Flickr8k benchmark to include more image-caption pairs, consisting of 30k images with 150k human-annotated captions, which are collected in the same way as Flickr8k.

The Microsoft Common Objects in Context (MS-COCO) dataset (Lin et al., 2014) is one of the most commonly used datasets to train and evaluate image captioning models. MS-COCO consists of 328k images collected from Flickr presenting real-world objects and scenarios. This dataset is created with extensive use of Amazon Mechanical Turk (AMT), with each distinct image paired with five human-annotated captions describing its visual contents.

Visual Genome (Krishna et al., 2017) is a another large-scale dataset for image captioning collected by crowd-sourcing. It consists of over 108k real-world non-iconic images from Flickr, each of which contains an average of 35 objects, 26 attributes and 21 pairwise relationships between objects. Visual Genome annotates separate captions for multiple regions in an image, providing a dense set of image descriptions.

VizWiz-Captions (Gurari et al., 2020) consists of 39K images originating from real-world use cases of captioning services for visually impaired people, while SentiCap (Mathews et al., 2016) provides several thousand images and accompanying captions with positive and negative sentiments.

Recently, there has been an increasing interest in using synthetic datasets as diagnostic

tools for image captioning models, with examples including CLEVR (Johnson et al., 2017) and ShapeWorld (Kuhnle and Copestake, 2017). The primary motivation behind developing these synthetic benchmarks is to reduce complexity that is considered irrelevant to the evaluation focus, and to enable better control over the content of training and testing data. These datasets can be highly valuable for evaluating image captioning models under controlled conditions with known ground truths, allowing researchers to identify areas for improvement and to explore the impact of specific model architectures and features on performance.

5.2.1.2 Existing models

Early approaches in image captioning addressed the task using a retrieval-based method (Ordonez et al., 2011; Gupta et al., 2012; Patterson et al., 2014). When presented with a query image, these models search for similar images in a database. A caption is then constructed based on the descriptions corresponding to the retrieved set of similar images. The caption for the query image can be generated by simply reusing the caption of the most similar image from the retrieved set, or by crafting a new description drawing from the descriptions of the similar images.

As deep learning approaches developed, they have increasingly been proposed for use in image captioning. Many of these models (Vinyals et al., 2015; Xu et al., 2015; Donahue et al., 2015; Yao et al., 2017c; Aneja et al., 2018) are built upon the typical end-to-end encoder-decoder framework, which is inspired by the neural machine translation framework (Sutskever et al., 2014; Bahdanau et al., 2014). In this framework, image captioning is treated as a translation task from images to natural language descriptions. These models learn to encode a source image into a static representation using convolutional neural networks (CNNs) and to decode a target caption from the image representation using recurrent neural networks (RNNs). In practice, long short-term memory models (LSTMs) (Hochreiter and Schmidhuber, 1997) are often used as the decoder given their ability to retain short-term memory over a relatively long period, alongside their performance in sequence tasks. The end-to-end design of neural models guarantees that all the parameters of a neural model can be jointly learned from the training data, alleviating the user from the burden of extensive feature engineering.

Building on the encoder-decoder architectural paradigm, various attention mechanisms have been adopted in recent work to enhance the performance of neural image captioning models. For example, Xu et al. (2015) integrated a hard stochastic attention mechanism and a soft deterministic attention mechanism to the encoder-decoder framework to decide which parts of the image should be attended to at each time step. You et al. (2016) proposed a neural image captioning model with semantic attention that detects semantic concepts as candidates for attention using a bottom-up approach, in addition to deciding

when and where to activate the attention according to a top-down visual feature. This guides the model to attend to semantically important regions or objects whilst encoding the images, thus preventing the model from being distracted by indifferent objects in the background. More recently, Anderson et al. (2018) further combine bottom-up and top-down attentions to enrich the visual feature information extracted from images, which has proven beneficial for both image captioning and visual question answering (VQA).

Whilst most image captioning models focus on the generation of captions in general, some studies have aimed to tackle more specific and challenging aspects of image captioning. These areas of investigation include generating captions for rare or unseen objects in visual scenes (Anderson et al., 2017; Yao et al., 2017b; Demirel and Cinbis, 2022), handling complex scenes containing multiple objects (Venugopalan et al., 2017), and producing captions that are more diverse and creative (Aneja et al., 2019; Chen et al., 2015; Khare and Huber, 2019; Padmakumar and He, 2022).

Inspired by the prevalent success of pre-trained NLP models such as BERT (Devlin et al., 2019), other recent studies have investigated vision-language pre-training approaches (Zhou et al., 2020; Kim et al., 2021; Yang et al., 2021; Liu et al., 2021a; Changpinyo et al., 2021; Hu et al., 2022). These pre-trained vision-language models are usually trained on large-scale image-text corpora to learn universal multimodal representations, which can be further fine-tuned on various downstream vision-language tasks. The most commonly used pre-training tasks for such training are arguably multimodal masked language modelling (Yang et al., 2021), a multimodal counterpart of the masked language modelling (MLM) task in the BERT model (Devlin et al., 2019). Other pre-training tasks such as image-text matching and multimodal contrastive learning are also widely used (Yang et al., 2021). Utilising these pre-trained multimodal representations has been reported to have achieved promising results for the downstream image captioning task, as well as for a variety of other multimodal tasks (Zhou et al., 2020; Kim et al., 2021; Hu et al., 2022).

5.2.1.3 Existing evaluation metrics

In addition to the automatic metrics that are widely used for most NLG tasks (e.g., BLEU, METEOR, etc), there are a number of metrics that are specially designed to evaluate image captioning models. SPICE (Anderson et al., 2016) is a recent evaluation metric specifically designed for the task of image captioning. SPICE parses both the candidate caption and reference captions to scene graphs, and then examines the agreement between logical tuples extracted from these scene graphs. The SPICE metric with its focus on semantic propositional content most closely relates to our faithfulness evaluation, although it still approaches propositional analysis as overlap comparison to reference captions.

An alternative approach used for image captioning evaluation proposes a ranking task for plausible alternatives (Hodosh et al., 2013). This approach offers the possibility

for adversarial evaluation where decoy captions are chosen based on their similarity to the target caption (Ding et al., 2016), or automatically generated as minimally different adversaries (Hodosh and Hockenmaier, 2016; Shekhar et al., 2017). However, these approaches are ultimately constrained by the quality of the images and captions, which are usually obtained from crowd-sourcing.

Recent work has further explored increasing the diversity of generated captions, for which various measures have been proposed. Devlin et al. (2015) investigates the concept of caption diversity by evaluating performance on compositionally novel images. van Miltenburg et al. (2018) frames image captioning as a word recall task and proposed several metrics, predominantly focusing on diversity at the word level. However, this direction is still relatively new and lacks standardised benchmarks and metrics.

5.3 The ShapeWorldICE benchmark

In this thesis, I constructed a synthetic benchmark designed for diagnostic image captioning evaluation. This evaluation benchmark is named ShapeWorldICE (*ShapeWorld for Image Captioning Evaluation*). Before diving into the ShapeWorldICE benchmark, I will first introduce the ShapeWorld framework (Kuhnle and Copestake, 2017) upon which this new benchmark is constructed.

5.3.1 The ShapeWorld framework

Recently, a multitude of synthetic datasets have been proposed as diagnostic tools for NLP models, such as CLEVR (Johnson et al., 2017) for visual question answering (VQA), the bAbI tasks (Weston et al., 2015) for text understanding and reasoning, and ShapeWorld (Kuhnle and Copestake, 2017) for visually grounded language understanding. The primary motivation behind building artificial data in these projects is to reduce visual or language (or both) complexity considered irrelevant to the evaluation focus, to enable better control over the content of training and testing data, and to provide more detailed insights into strengths and limitations of learned model behaviours. van Miltenburg et al. (2020) advocated the use of synthetic data which enables practitioners to learn more about the models that they build for real-world applications. As one can control the level of cleanness (or messiness) of the synthetic data that is curated, the evaluation for models trained and tested on such data is likely to be more informative about what a model can or cannot handle.

In this work, I developed an image captioning evaluation benchmark utilising the ShapeWorld framework (Kuhnle and Copestake, 2017). ShapeWorld is a controlled data generation framework consisting of abstract coloured shapes (see Figure 5.2 for an example). ShapeWorld has proven to be useful for in-depth inspection of VQA-style models (Kuhnle

Type	Variant	Caption	Image
Exist	OneShape	There is a green cross. A rectangle is green. There is a cyan shape.	
	MultiShapes	A shape is a gray triangle. There is a square. There is a yellow circle.	
Spatial	TwoShapes	The lowermost red shape is a pentagon. There is a cross above a yellow shape. A yellow square is to the left of a pentagon.	
	MultiShapes	A triangle is to the left of a semicircle. A circle is above a green rectangle. A semicircle is to the left of a circle.	
Quant	Count	Exactly one circle is yellow. More than one of the green shapes are rectangles. Exactly zero shapes are ellipses.	
	Ratio	A quarter of the shapes are rectangles. A third of the rectangles are magenta squares. At least half the shapes are green rectangles.	

Table 5.1: Examples from the ShapeWorldICE datasets (faithful captions in green, false in red). Images from `Existential-OneShape` contain one object, while images from `Spatial-TwoShapes` contain two objects. Images from the other four datasets follow the same distribution with multiple abstract objects present in a visual scene.

et al., 2018). In this thesis, ShapeWorld is used to generate training and evaluation data for two major reasons. Firstly, ShapeWorld supports customised data generation according to user specification, which enables a variety of model inspections in terms of language construction, visual complexity and reasoning ability. The second benefit is that each training and test instance generated in ShapeWorld is returned as a triplet of <image, caption, world model>. The **world model** stores information about the underlying micro-world used to generate an image, internally represented as a list of entities with their attributes, such as *shape*, *colour* and *position*. During data generation, ShapeWorld randomly samples a world model from a set of available entities and attributes. The generated world model is then used to realise a corresponding instance consisting of an image and a caption. Each image is a direct visualisation based on the attribute values contained in the sampled world model. On the linguistic side, ShapeWorld utilises deep semantic representations based on dependency minimal recursion semantics (Copestake, 2009, DMRS) and the broad-coverage English Resource Grammar (Flickinger, 2000, ERG) to realise an underlying world model representation to a textual caption. As such, every image rendered within ShapeWorld is a straightforward visualisation of a particular world model, whilst every caption is also directly realised from the world model. The world model provides the actual semantic information contained in an image, which allows the

evaluation of faithfulness of a caption to a particular image.

5.3.2 ShapeWorldICE for image captioning evaluation

The present study introduces the ShapeWorldICE benchmark, which comprises synthetic data specifically tailored for the task of image captioning using the ShapeWorld data generation framework (Kuhnle and Copestake, 2017). Expanding upon the original ShapeWorld framework, three distinct types of ShapeWorldICE datasets were constructed, including `Existential`, `Spatial`, and `Quantification`. Each dataset emphasises a specific aspect of the captioning task. `Existential` captions focus on determining the presence of objects within an image, while `Spatial` captions describe the fundamental spatial relationships between objects in a visual scene. On the other hand, `Quantification` captions provide count-based or ratio-based descriptions of objects within an image. Moreover, two variants were developed for each caption type, emphasising different levels of visual complexity or specific reasoning aspects. For example, the `Existential-OneShape` (denoted as `Exist-OneShape`) images consist of a single coloured object in an image, whilst the `Existential-MultiShapes` (abbreviated as `Exist-MultiShape`) images contain multiple objects in a visual scene. The `Spatial` images also contain two variants. The first variant, `Spatial-TwoShapes`, consists of images containing only two objects. The corresponding captions for these images aims to describe the spatial relationship between the two objects. The second variant, `Spatial-MultiShapes`, pertains to images with multiple objects, typically exceeding two. In this case, the captions focus on spatial relationships involving two of the objects within the multi-object image. The `Quantification-Count` captions, referred to as `Quant-Count`, concern the number of specific types of objects within a multi-object image. Conversely, the `Quantification-Ratio` captions, referred to as `Quant-Ratio`, approach the quantification problem from a different perspective, describing the proportion of a particular type of objects relative to the total number of objects in the image. A comprehensive overview of the ShapeWorldICE datasets developed in this work is presented in Table 5.1.

In accordance with standard machine learning practices, every ShapeWorldICE dataset is partitioned into three splits: a training set containing 200,000 instances, a validation set containing 4,096 instances, and a test set containing 4,096 instances. Each training and validation instance comprises an image and a corresponding reference caption. Note that the GFD framework is a **reference-less** evaluation framework, meaning that it does not rely on reference captions to carry out evaluation on the test set. However, to enable the comparison of our proposed GFD framework with reference-dependent metrics such as BLEU (Papineni et al., 2002) and SPICE (Anderson et al., 2016), I randomly sampled ten reference captions for each instance in the test set.

5.4 Evaluating grammaticality for generated captions

An essential requirement for an image captioning system to be considered as competent is to be able to produce captions that are grammatically well-formed. Assessing grammaticality for a general piece of text is a difficult task in itself, but becomes more feasible within a constrained context such as the synthetic ShapeWorldICE benchmark.

5.4.1 Parsability with the ERG as a proxy

As introduced in chapter 3, the **parsability** of a model-generated caption with the ERG (Flickinger, 2000) is taken as a *surrogate* for the evaluation of grammaticality. In this sense, a candidate caption is considered as grammatical if it is possible to obtain a valid parse using the ERG. Given a set of N captions generated by an image captioning model, the grammaticality score (denoted as G) obtained by the model on the test samples can be formally defined as:

$$G = \frac{\#\{parsable\}}{N}$$

where $\#\{parsable\}$ denotes the number of captions that can be successfully parsed with the ERG.

5.4.2 Baseline models

I conducted experiments with two contemporaneously proposed image captioning models: the Show&Tell model (Vinyals et al., 2015) and the LRCN_{1u} model (Donahue et al., 2015). Both models follow the typical architecture used in image captioning tasks. They employ a CNN-based component to encode images into feature representations, and an LSTM-based component to decode a natural language sequence based on the image representations. The main difference between these two models lies in how they handle inputs during the caption generation process. In the Show&Tell model, an image representation is only provided once at the beginning of the decoding process to inform the decoder about the content of an image, whilst the LRCN_{1u} model feeds the image representation along with the embedded representation of the previously generated word to the decoder at each time step during the caption generation process.

In this work, the Show&Tell model was implemented using the open-sourced im2txt library[2]. The LRCN_{1u} model is a TensorFlow implementation of the initial Caffe code[3]. Word embeddings in the LSTM were randomly initialised, whilst initial weights in the CNN component were obtained by pre-training the component on a relevant object detection task.

[2] github.com/tensorflow/models/tree/master/research/im2txt
[3] github.com/jeffdonahue/caffe/tree/recurrent/examples/coco caption

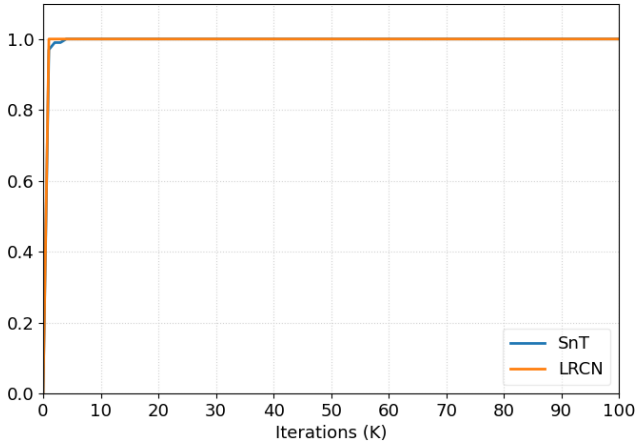


Figure 5.3: Evaluation results of grammaticality for the Show&Tell model and the LRCN_{1u} model on `Exist-MultiShapes`. *SnT* denotes the Show&Tell model, whilst *LRCN* denotes the LRCN_{1u} model.

Following established best practice in transfer learning, I fine-tuned the models on the ShapeWorldICE datasets. Models are trained and tested in an end-to-end process. Each training instance consists of an image and a reference caption. During fine-tuning, the CNN encoder and the LSTM decoder of both models are jointly trained by optimising the cross-entropy training loss. I used the Adam optimiser (Kingma and Ba, 2014) with a learning rate of 0.001 and a batch size of 64. I fine-tuned model hyperparameters based on the performance on the validation set. At test time, only the test images were available to the evaluated models. Underlying world models were kept from the models, and were only used for later GFD evaluation.

5.4.3 Results

Model comparison. Figure 5.3 presents the grammaticality scores obtained by Show&Tell and LRCN_{1u} on the test set of `Exist-MultiShapes`. All reported results were measured on the test split, employing the model parameters that yielded superior validation performance. To investigate the progression of model performance throughout the training process, the grammaticality scores obtained by each model on the test set were systematically recorded. This recording took place as the models underwent training for a fixed number of 100,000 iterations. The observed trend reveals that both models exhibit an early ability to produce grammatically well-formed captions, even from the initial stages of training.

Model inspection. I further examined the ratios of grammatical sentences produced by the LRCN_{1u} model for different types of ShapeWorldICE datasets, including `Exist-OneShape`, `Exist-MultiShape`, `Spatial-TwoShapes`, `Spatial-MultiShapes`, `Quant-Count` and `Quant-Ratio`. The objective was to investigate whether language structure variations in captions across different datasets impacted the grammaticality of the

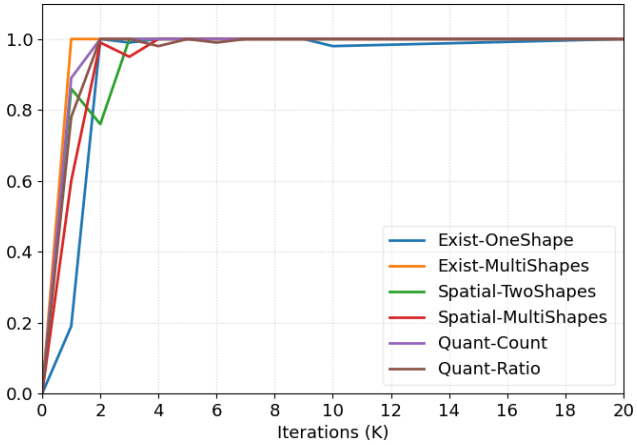


Figure 5.4: Grammaticality scores for LRCN_{1u} on `Exist-OneShape`, `Exist-MultiShapes`, `Spatial-TwoShapes`, `Spatial-MultiShapes`, `Quant-Count` and `Quant-Ratio`. Results are cropped to only report the first 20,000 iterations, as the grammaticality scores stay at 1.0 afterwards for all ShapeWorldICE datasets.

model's generated descriptions.

Figure 5.4 presents the curves of grammaticality scores obtained by LRCN_{1u} on different ShapeWorldICE datasets during the first 20,000 training iterations. The Show&Tell model exhibited a similar trend. Notably, the graph indicates that the LRCN_{1u} model consistently achieved near-perfect grammaticality scores for captions generated across all ShapeWorldICE data types in under 5,000 training iterations, showcasing the model's rapid acquisition of the ability to generate grammatically well-formed sentences.

It is important to mention that all training captions in ShapeWorldICE are meticulously rendered using the ERG in a controlled manner. Consequently, it is unsurprising that statistical models trained on the ShapeWorldICE data can generate captions that successfully pass the parsing tests within the same grammar engine, attaining correspondingly high grammaticality scores.

5.5 Evaluating faithfulness for generated captions

The combination of a fully controlled data configuration and full access to underlying world models in ShapeWorldICE enable a thorough examination of the consistency between a caption and the actual visual content the caption aims to describe.

5.5.1 Evaluating faithfulness against world models

The evaluation of caption faithfulness relies heavily on the underlying world models, as they encode the ground truth attributes and values of specific visual scenes.

In this work, I adopted a deep-semantics inspired approach leveraging the DMRS representation (Copestake, 2009), which was initially used in the ShapeWorld framework to render captions from world models. During the faithfulness evaluation process, a caption generated by a model is first parsed with the ERG, following a similar procedure to the aforementioned grammaticality evaluation. The resulting ERG parse is converted into a DMRS graph using the `pydmrs` tool (Copestake et al., 2016). As such, each DMRS graph serves as a logical semantic graph representation of the original caption. I then construct a logical predicate based on the converted DMRS graph and evaluate it against the underlying world model associated with the corresponding image. By examining the logical agreement between the logical predicate representation of a caption and the world model of an image, it is possible to assess the agreement of the caption to its corresponding image.

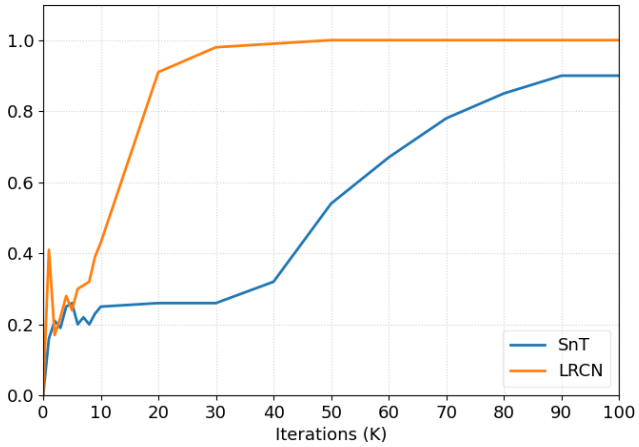
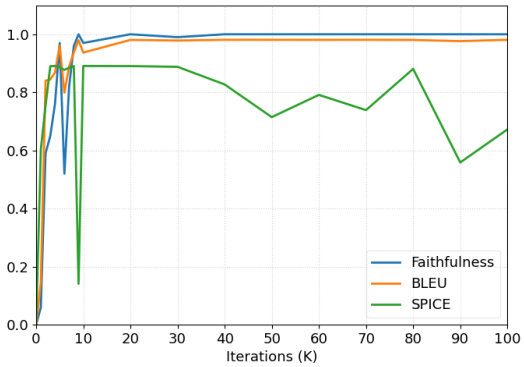


Figure 5.5: Evaluation results of faithfulness for the Show&Tell model and the LRCN_{1u} model on `Exist-MultiShapes`. *SnT* denotes the Show&Tell model, whilst *LRCN* denotes the LRCN_{1u} model.

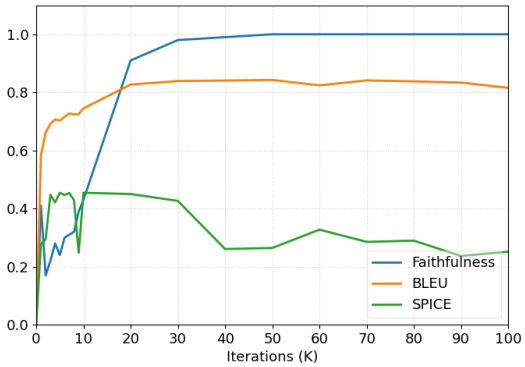
5.5.2 Results

Model comparison. Faithfulness results for the Show&Tell model and the LRCN_{1u} model on `Exist-MultiShapes` are presented in figure 5.5. It can be seen that LRCN_{1u} is clearly superior in terms of faithfulness, achieving 100% faithfulness halfway through training, whereas Show&Tell only slowly attains around 90% faithfulness by the end of 100,000 iterations. This indicates that incorporating visual features at every generation step is beneficial for producing faithful captions. I observed similar results with other ShapeWorldICE datasets that I experimented with, validating the superiority of LRCN1u over Show&Tell in terms of generating faithful captions on ShapeWorldICE.

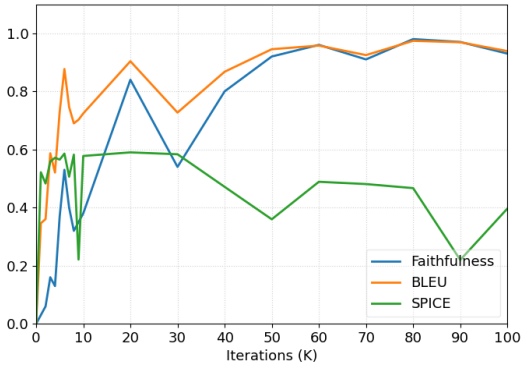
Correlation with BLEU/SPICE scores. The analysis of the results presented in figure 5.6 highlights a noticeable lack of correlation between BLEU/SPICE scores and



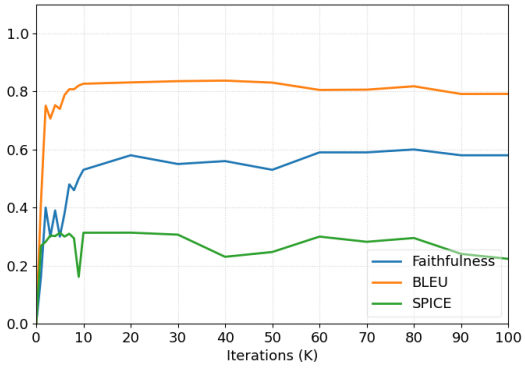
(a) Exist-OneShape



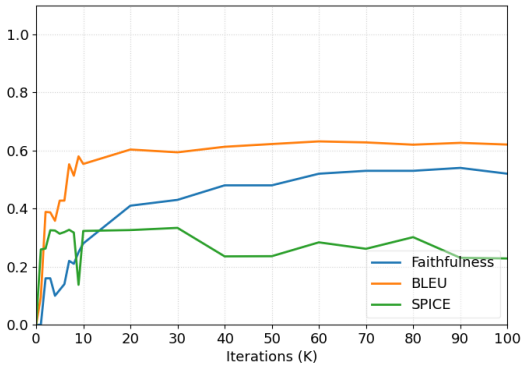
(b) Exist-MultiShapes



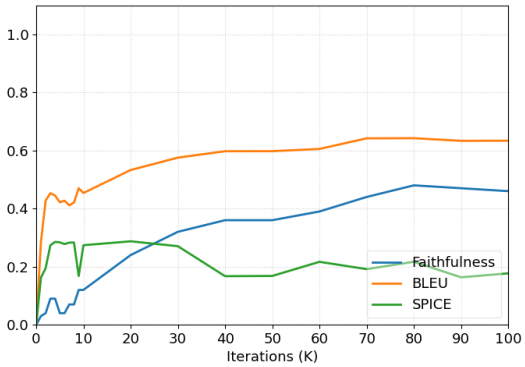
(c) Spatial-TwoShapes



(d) Spatial-MultiShapes



(e) Quant-Count



(f) Quant-Ratio

Figure 5.6: Evaluation results with different metrics for LRCN_{1u} on `Exist-OneShape`, `Exist-MultiShapes`, `Spatial-TwoShapes`, `Spatial-MultiShapes`, `Quant-Count` and `Quant-Ratio`. *Faithfulness* refers to the faithfulness score, i.e., the ratio of generated captions that agree with ground-truth world models. *BLEU* and *SPICE* denote the average BLEU-4 score and the average SPICE score obtained across the test split, respectively.

caption faithfulness. While the BLEU score seems to be a reliable indicator of caption faithfulness in the simple scenario of `Exist-OneShape`, its effectiveness diminishes in more complex scenarios. Specifically, in the case of `Spatial-MultiShapes`, spatial descriptors are chosen from a fixed set including "*above*", "*below*", "*to the left of*" and "*to the right of*". In this case, there is a high likelihood that a generated spatial descriptor matches one

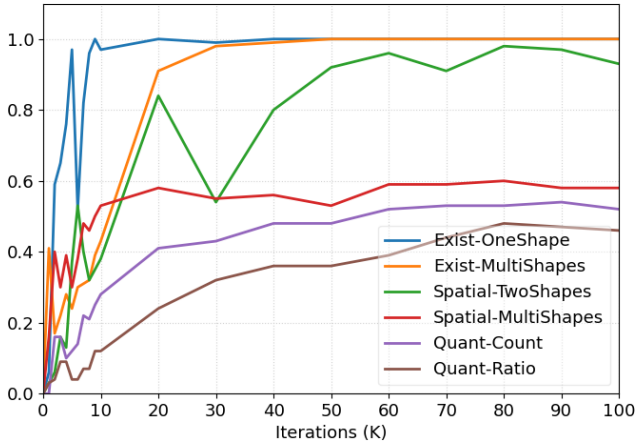


Figure 5.7: Faithfulness scores for LRCN_{1u} on `Exist-OneShape`, `Exist-MultiShapes`, `Spatial-TwoShapes`, `Spatial-MultiShapes`, `Quant-Count` and `Quant-Ratio`.

of the descriptors in the reference captions, leading to inflated BLEU scores. As such, it is evident that an increased BLEU score does not necessarily imply improved performance, thereby underscoring the limitations of BLEU as a comprehensive measure of caption faithfulness.

Despite SPICE's consideration of the semantic aspects of a visual scene to a greater extent than n-gram-based metrics, it still faces an inherent and systematic challenge due to its reliance on surface reference captions as a proxy for the actual visual contents. Figure 5.6a demonstrates this issue, as the SPICE curve exhibits a downward trend in the later stage of training on the `Exist-OneShape` dataset, while both the faithfulness score and BLEU score show rapid early improvement followed by plateauing into a high score afterwards. SPICE is calculated as the F1 score of scene graph matching between the candidate and reference scene graphs. In cases where multiple objects are present in an image, there can be multiple valid descriptions that are not referred to in the reference captions. Consequently, a caption that overlaps only partially with the scene graph will be penalised due to an imperfect recall score, regardless of its actual agreement with the image. This issue highlights the limitation of SPICE in accurately capturing the faithfulness of captions, particularly when considering the full complexity and variability of visual scenes.

Model inspection. Similarly to the grammaticality evaluation, the faithfulness evaluation metric can also be used to inspect a specific model architecture. The closed-world controlled specification of ShapeWorldICE provides an ideal setting for a diagnostic evaluation of caption faithfulness, especially across varying training data scenarios. To investigate how visual complexity and linguistic construction affect the behaviour of an image captioning model, I trained the LRCN_{1u} model on various ShapeWorldICE datasets and subsequently generated faithfulness curves for these model variants over 100,000 training steps, as depicted in figure 5.7.

An important observation from the evaluation results in figure 5.7 is the model's inability to effectively learn complex spatial relationships. While caption faithfulness score remains relatively high in the simple `Spatial-TwoShapes` scenario, it significantly drops when dealing with more intricate visual scenes in the `Spatial-MultiShapes` dataset. This substantial decrease in faithfulness clearly highlights the model's deficiency regarding its ability to directly acquire spatial relationships from complex visual scenes.

Furthermore, the experiment results obtained from the count-based and ratio-based `Quantification` datasets display the poor performance of the LRCN_{1u} model in terms of caption faithfulness. Specifically, the model achieved faithfulness scores of 0.50 and 0.46 on the `Quant-Count` and `Quant-Ratio` tasks respectively. These findings indicate that counting-related tasks pose a significant challenge for the LRCN_{1u} architecture, further underscoring the limitations of the model in accurately capturing and expressing quantitative information in captions.

5.6 Evaluating diversity for generated captions

Recent studies have highlighted that maximum likelihood estimation (MLE) training encourages models to generate captions that closely resemble those in the training set. Consequently, models tend to generate uniform and generic descriptions for similar images (Dai et al., 2017; van Miltenburg et al., 2018). This proven deficiency provides the impetus for diversity to be established as the third fundamental criterion in caption evaluation.

It is worth noting that the initial diversity evaluation described in the paper "Going Beneath the Surface: Evaluating Image Captioning for Grammaticality, Truthfulness and Diversity" (Xie et al., 2020) was tailored to ShapeWorld[4], utilising the pre-defined construction rules available in the world models in ShapeWorld. In this thesis, I have focused on two other aspects of diversity evaluation: word-level diversity and parse-level diversity. These two aspects are more versatile and applicable to complex scenarios where the underlying construction rules may not be readily available. They will also be utilised in the diversity evaluation for the two upcoming tasks introduced in subsequent chapters.

5.6.1 Word-level diversity

Another level of diversity I consider is word-level diversity. The word-level diversity of a sentence is defined as the **normalised type-token ratio (TTR)** of the sentence.

[4]Given that ShapeWorld inherently exploits a pre-defined set of caption constructions, it is possible to measure the diversity of model-generated caption constructions by comparing them to those rendered in the training data by the ShapeWorld generation framework. To achieve this, the language constructions sampled in the training captions are taken as a proxy for optimal caption diversity. Further details on how to calculate the ShapeWorld-specific diversity score can be found in Xie et al. (2020).

The conventional version of TTR (Richards, 1987) calculates the ratio of the number of different *types* to the number of all *tokens* in a text:

$$\text{TTR} = \frac{\#type}{\#token} \tag{5.1}$$

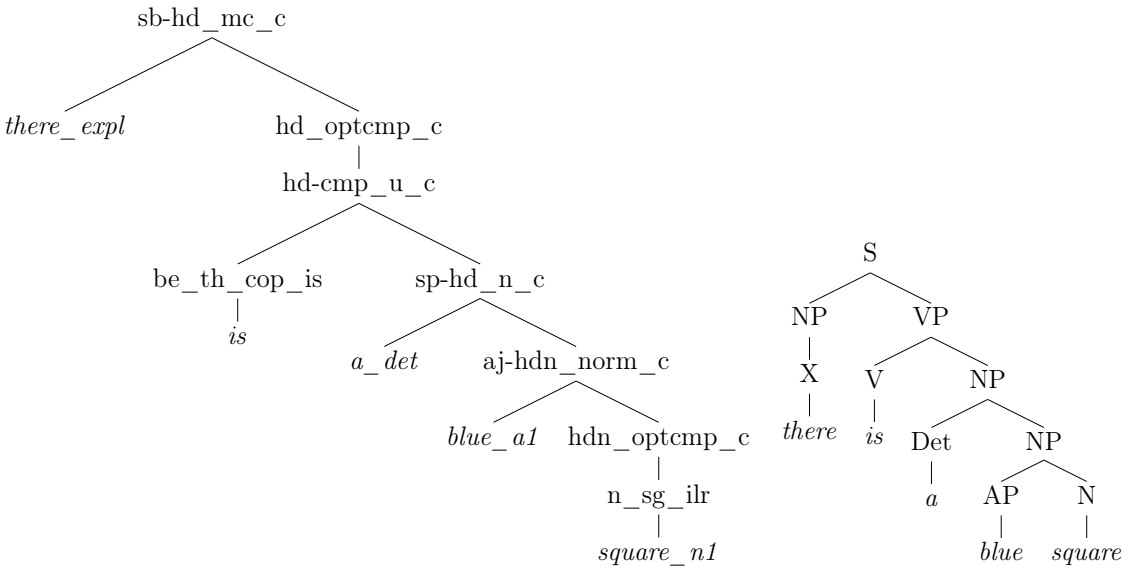
Conventional TTR's usefulness primary lies in revealing lexical richness when the texts to be compared are of the same length. A high TTR indicates a significant level of lexical variation; the larger the resulting TTR, the less repetitive the vocabulary use. However, when text lengths vary, the conventional TTR is negatively correlated with the number of tokens in a text (Richards, 1987). Long sentences tend to get rated below shorter sentences using the conventional TTR. This is understandable - the longer a text runs on, the fewer examples of novel vocabulary will be introduced. To address this issue, I implement an alternative variant of TTR with length normalisation - moving average type-token ratio (Covington and McFall, 2010, MATTR) - as the word-level diversity measure. MATTR works by computing a conventional TTR for every n running words (called a *window*) in a text, and then taking the average of the TTRs for all chunks as the normalised TTR. Formally, given a set of model predictions $y_1, y_2, \ldots, y_N$ and a pre-set window n, all tokens in the concatenated sequence $[y_1; y_2; \ldots; y_N]$ can be sliced into chunks of tokens of the window size n. Denote the resulting chunks as $T_1, T_2, \ldots, T_m$, where T_i is the i-th trunk containing n tokens, and m is the number of all sliced chunks. The word-level diversity (denoted as DIV_{wl}) is defined as

$$\begin{aligned} DIV_{wl} &= \text{MATTR}(y_1, \ldots, y_N) \\ &= \frac{\sum_{i=1}^{m} \text{TTR}(T_i)}{m} \end{aligned} \tag{5.2}$$

When computing the number of *types* in a text, there are multiple linguistic elements that can potentially be used as the *type* of a word. In this thesis, I used the base dictionary form of a word (known as a **lemma**) for the categorisation of *types*. The reason of choosing lemmas over other linguistic elements (such as stems) is that the lemmatisation process takes into account the morphological analysis of a word to find its basic dictionary form. As a result, a lemma is the basic form of all possible inflectional forms, whereas a stem is not. As such, using lemmas is a more sensible choice concerning representing *types*. I used the NLTK implementation of the WordNetLemmatizer (Bird et al., 2009) to extract lemmas from words.

5.6.2 Parse-level diversity

Whilst word-level diversity examines lexical difference, parse-level diversity focuses on structural variety within model outputs. Parse-level diversity examines the number of



(a) Syntactic derivation tree. (b) Phrase-structure tree.

Figure 5.8: Syntactic derivation and phrase structure trees for "*There is a blue square*".

different constructions a model can generate, and the construction variety between model outputs. In parse-level diversity evaluation, the model outputs are first parsed into derivation trees in which key information about grammar construction in the input is enclosed. From the derivation trees, we can obtain a set of grammar rules that occur in the model outputs. The outputs are then mapped to vector representations based on the grammar rules that are present in individual parse trees. This vectorisation step converts a natural language sentence into an embedding which represents grammar rule presence in the original sentence. Parse-level diversity is calculated as the sparsity of the vector cluster, weighted by the number of different constructions seen in this output group.

5.6.2.1 Derivation trees

A derivation tree, or a parse tree, is an ordered, rooted tree that represents the syntactic structure of a sentence. Derivation trees encode important syntactic information about linguistic inputs. In this work, I used the DELPH-IN syntactic derivation tree (Ivanova et al., 2012, 2013) as an instance of the derivation trees. The syntactic derivation tree is a head-driven phrase structure grammar (HPSG) derivation tree. The rationale behind choosing the syntactic derivation tree over other representations is that the syntactic derivation structure provides fine-grained, complete information to rebuild a sentence or replay an HPSG analysis. Figure 5.8 compares the syntactic derivation tree of the sentence "*There is a blue square*" with its phrase structure analysis. The syntactic derivation tree differs from the traditional phrase structure tree mostly in the granularity and the level of hierarchy encoded in their representations. Traditional phrase structure trees usually provide information about lexical nodes and shallow grammar entries (labelled with

"S", "NP", "VP", etc). In contrast, the syntactic derivation trees not only encode *lexical information*, but also present hierarchical *construction rules* of syntactic components.

A syntactic derivation tree is composed of identifiers for grammar entities, including grammar rules and lexical entries. There are three types of nodes in a syntactic derivation tree: the *root node* which denotes the abstract root of a parse, the *phrasal nodes* which store information about grammar rules, and the *leaf nodes* which contain the input tokens and their fine-grained lexical information. For the syntactic derivation tree illustrated in figure 5.8a, phrasal nodes denote grammar rules exhibited in the sentence, such as *sb-hd_mc_c* (head + subject, main clause), *hd-cmp_u_c* (head + complement), and *aj-hdn_norm_c* (nominal head + preceding adjunct). Leaf nodes denote lexical entries in the input, such as *a_det* ("a"), *blue_a1* ("blue"), and *square_n1* ("square").

For the calculation of parse-level diversity, the *phrasal nodes* are most relevant as they contain the entity names of specific grammar rules and the locations where the rules are applied.

5.6.2.2 Vectorisation

At the vectorisation step, model-generated sentences are mapped to vector representations. The mapping is performed based on whether a specific grammar rule is present in a highlight's derivation tree.

Suppose a generation model outputs a number of sentences $y_1, y_2, \ldots, y_V$ for the test set containing V test samples. The derivation trees parsed from the outputs are $T_1, T_2, \ldots, T_V$. By iterating over the derivation trees and finding all the phrasal nodes, we can obtain a set of unique grammar rules that appeared in the nodes, denoted as $r_1, r_2, \ldots, r_D$.

For each model-generated caption y_i, we initialise a D-dimensional vector $\mathbf{v}^i$. Each component in $\mathbf{v}^i$ is decided as below:

$$\mathbf{v}^i_j = \begin{cases} 1, & \text{if } r_j \text{ in } T_i \\ 0, & \text{otherwise} \end{cases} \tag{5.3}$$

where $j \in [1, 2, \ldots, D]$.

The resulting vector $\mathbf{v}^i$ of a caption y_i is a D-dimensional vector consisting of 0s and 1s as its components. The value of each vector component indicates the presence of a specific grammar rule in the caption y_i. After the mapping for all model-generated outputs, we have a cluster of V vectors with dimension D.

5.6.2.3 Parse-level diversity

To calculate how diverse a set of model outputs are with respect to grammatical constructions, we can examine how compact the vectors obtained from the outputs using equation

5.3 are, and how many constructions the model can generate. The more compact the obtained vectors are, the more similar the grammatical constructions of the model outputs are.

I first calculate the *sparsity* of the group of grammar rule vectors which are obtained from the model outputs. The sparsity of a group of vectors is calculated as the **mean square distance** of the vectors to the centroid vector (the average of all vectors). Here the **Euclidean distance** between vectors is used as the geometric distance metric.

$$Sparsity(\mathbf{v}^1, \ldots, \mathbf{v}^V) = \frac{1}{V}\sum_{i=1}^{V}(dist_{Euc}(\mathbf{v}^i, \overline{\mathbf{v}}))^2 \tag{5.4}$$

where $\overline{\mathbf{v}}$ is the centroid vector of $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^V$, and $dist_{Euc}(\mathbf{v}^i, \overline{\mathbf{v}})$ is the Euclidean distance of vectors $\mathbf{v}^i$ and $\overline{\mathbf{v}}$:

$$dist_{Euc}(\mathbf{v}^i, \overline{\mathbf{v}}) = \sqrt{\sum_{j=1}^{D}(\mathbf{v}^i_j - \overline{\mathbf{v}}_j)^2} \tag{5.5}$$

where D is the dimension of vector $\mathbf{v}^i$.

Recall that the *variance* describes the level of variability around the expectation and is calculated as the expected value of the squared difference between every element and its expected value. Therefore, $Sparsity(\mathbf{v}^1, \ldots, \mathbf{v}^V)$ is equivalent to the variance of the Euclidean distance of all vectors in the cluster.

The *parse-level diversity* (denoted as DIV_{pl}) of the outputs of a model is calculated as follows:

$$DIV_{pl} = Sparsity(\mathbf{v}^1, \ldots, \mathbf{v}^V) * log(D) \tag{5.6}$$

DIV_{pl} reflects a model's ability of generating structurally different candidates. The bigger the value DIV_{pl} is, the higher parse-level diversity the outputs of a model have.

5.6.3 Results

Model inspection. Table 5.2 presents the word-level and parse-level diversity scores obtained by LRCN_{1u} on the full range of datasets. The evidence suggests that the diversity of inferred captions is largely sensitive to the inherent variability of captions within each dataset. In the case of simpler datasets such as `Exist-OneShape`, the model exhibits a tendency to adopt uniform sentence structures when generating captions, as illustrated by the low parse-level diversity score of 0.289 in table 5.2 on this particular dataset. More complex datasets such as the two `Quant` variants clearly exhibit higher parse-level diversity compared to other dataset variants. From this it can be inferred that caption diversity benefits from the inclusion of the diverse and heterogeneous language constructions found in more complex datasets.

Type	Variant	DIV_{wl}	DIV_{pl}
Exist	OneShape	0.356	0.289
	MultiShapes	0.368	0.696
Spatial	TwoShapes	0.526	0.340
	MultiShapes	0.493	0.389
Quant	Count	0.527	2.809
	Ratio	0.499	3.645

Table 5.2: Word-level and parse-level diversity scores for the LRCN_{1u} model obtained on `Exist-OneShape`, `Exist-MultiShapes`, `Spatial-TwoShapes`, `Spatial-MultiShapes`, `Quant-Count` and `Quant-Ratio`. DIV_{wl} and DIV_{pl} refer to word-level and parse-level diversity scores, respectively. The word-level diversity scores were calculated with a sliding window size of 15 words.

5.7 Discussion

In this chapter, I introduced the first task scenario for the GFD evaluation framework, facilitated by the ShapeWorldICE benchmark. The synthetic environment provided by ShapeWorldICE played a crucial role in creating a controlled setting for a diagnostic evaluation of the models. This setup allowed for a detailed study of their behaviours across different levels of visual and linguistic complexity.

The in-depth examination of the GFD evaluation framework on two image captioning models illuminated the complex inner workings of these models, offering understanding into their performance based on various criteria. The grammaticality assessment, conducted with the ERG, revealed an early competency in both models to produce captions that are grammatically well-formed, regardless of the linguistic complexity of the training data. A contrast between the faithfulness evaluation and traditional metrics showcased the limitations of BLEU and SPICE as reliable measures in revealing caption faithfulness. Their dependency on comparisons with reference captions led to an insufficient capture of the genuine faithfulness of a generated caption to its corresponding world model. Finally, the exploration of caption diversity underscored that diversity is not only model-dependent but also reliant on the intrinsic variability of captions within each dataset. More complex datasets with diverse and heterogeneous language constructions promoted greater diversity in the generated captions.

The experimental findings in this chapter exposed the strengths and weaknesses of the two image captioning models, showcasing the effectiveness of the GFD evaluation framework in the context of this task. Additionally, this research highlighted the potential of synthetic data in facilitating diagnostic evaluations of image captioning models. The ShapeWorldICE benchmark, with its completely controlled data generation setup, provided an ideal platform for conducting fine-grained inspection of model performance.

Chapter 6

FHiG: Football highlight generation and evaluation

In chapter 5, I discussed the task of image captioning using synthetic data and its GFD evaluation framework. I will move on to another generation task - **F**ootball **H**ighl**i**ght **G**eneration (FHiG). I consider this task an appropriate next step for the ShapeWorldICE evaluation work. As discussed in chapter 5, ShapeWorldICE's synthetic scenes enable a finely controlled evaluation which could shed light on models' strengths and limitations. In a synthetic setting, the visual ground truth is strictly controlled. It is therefore possible to perform a very fine-grained, diagnostic evaluation for aspects that are of particular importance for us. This is especially true for the faithfulness evaluation, where access to the ground truth is essential. However, this synthetic setup limits its use in real-world scenarios. The FHiG task bridges the requirements of a controlled generation setup and a real-world scenario, due to the semi-structured nature of its input.

In the subsequent sections, I will first give a formal definition of the task in section 6.1. I will then introduce the dataset that is specifically curated for the purpose of this task in section 6.2. Similar to the ShapeWorldICE work, a GFD evaluation framework is proposed for the football highlight generation task. Baseline model implementation and experiment results are presented in sections 6.3-6.7, followed by a discussion in section 6.8.

6.1 The task of football highlight generation

The task of football highlight generation (FHiG) falls into the broader research area of data-to-text generation. In a data-to-text scenario, generation systems are expected to predict an output y in a natural language form given an input x, which is usually in the form of semi-structured data (such as diagrams and tables). FHiG is an instance of the data-to-text task, in that the input of FHiG is *match statistics* and the output is a *match*

highlight. Here a *match highlight* is a natural language description of an important event in a football game, usually containing 1-3 natural sentences. Note that in this chapter I will use the phrase *match result* and *match statistics* interchangeably - they both refer to the outcome of a football match.

Given an event x in a football match, we denote $W(x)$ as the set that contains all the information available for x. $W(x)$ can be considered as the underlying world model for the event x.

Define $I(x)$ as the information that is used to generate a sequence $y = y_1, y_2, \ldots, y_M$ that describes the event x. From the definition of $I(x)$, we have $I(x) \subseteq W(x)$. We refer to y as a *highlight* of a football match as the event x is usually chosen from a list of noteworthy events during a match. The task of football highlight generation can then be formalised as:

$$y = G(I(x)) \tag{6.1}$$

where G is the generation function. I will present a few approaches that are used as the generation function for this task in section 6.3.

The world model $W(x)$ can be represented as a set of attributes associated with the event x:

$$W(x) = e_1 : v_1, e_2 : v_2, \ldots, e_N : v_N \tag{6.2}$$

where e_i is an attribute (e.g., "event type") and v_i is the value of the attribute e_i.

There are two types of attributes that are associated with a particular event x. I use **primary attributes** to refer to the attributes that are event-specific and are different for events of the same match. Among the primary attributes for an event x, a key attribute is the *event type*. I focus on four major event types during a football match: goal, red card, half time and full time.

- **Goal**. The primary objective of a football game is to score goals. When a player successfully puts the ball into the opposing team's net, it results in a goal.
- **Red card**. During a football game, when a player commits a malicious offence, the referee has the authority to issue a caution in the form of a yellow card. If the offence is severe or if the player persistently violates the rules, the referee may issue a red card. A red card results in the player being expelled from the field, leaving their team with a numerical disadvantage.
- **Half time**. Half time, also referred to as the halftime interval or simply "the break", is the mid-game period between the two halves of a football match. It takes place midway through the game, typically after 45 minutes of play.
- **Full time**. Full time refers to the completion of a football match, which consists of two halves of 45 minutes each. Additional stoppage time may be added by the

referee to compensate for interruptions. After the completion of the second half and any additional time, the referee blows the final whistle, signifying the end of the match. The final score and outcome of the game are determined at full time.

These four events are specifically chosen as they typically showcase the most significant and exciting moments of a football match. Goals are often the most celebrated and captivating moments in a football match, as a well-executed goal demonstrates the offensive prowess, teamwork, and striking ability of the scorers. Additionally, the issuance of a red card by the referee for serious offences or repeated rule violations results in the player being sent off the field, creating a numerical disadvantage for their team. This can significantly impact the direction and progression of the game. As a red card introduces more uncertainty and excitement to a football game, it has the potential to be a pivotal turning point of the match. In football games, half time scores and full time scores are typically announced to provide updates on the progress and outcome of a match. The announcements of these scores allow spectators to stay informed about the current state of the game. They provide valuable information about the performance of the teams and contribute to the overall experience and understanding of the match. Given their significance, these four event types are well-suited to be included in a match highlight as they encompass pivotal moments that have a significant impact on the outcome of the game.

The other primary attributes of an event x are dependent on the value of their *event type*. I list below the other primary attributes that are collected for each *event type*:

Goal: the *time* of the goal, the *player* who scores the goal, and the *team* that the player plays for.

Red card: the name of the *player* who gets the red card, and which *team* the footballer is playing for.

Half time: the *scores* of both teams at half time.

Full time: The *scores* of both teams at full time, and the final outcome. If the match does not end in a draw, the *winning team* is also recorded.

Attributes that are shared across all events in the same football match are referred to as **secondary attributes**. Some examples of the secondary attributes are: the *name of a stadium* that a football game is held and its *location*, the *kickoff time*, the *name of the home team*, and the *name of the away team*.

6.1.1 Related work in data-to-text generation

Data-to-text generation refers to the task of generating natural language descriptions conditioned on structured input data (Reiter and Dale, 2000). This task can be considered as the most traditional form of NLG, which refers to the process of generation information in linguistic form from some underlying representation of information. The underlying representations are typically structured non-linguistic data, often in the form of tables, graphs, and knowledge bases.

Tasks and datasets. Early data-to-text tasks focused primarily on the generation of textual summaries derived from database records. Alongside text generation, systems for these tasks also conducted data analysis. A notable example of such generation tasks is weather forecasting (Goldberg et al., 1994; Reiter et al., 2005), where meteorological data is transformed into detailed weather reports. Automated journalism, also known as robo-journalism (Kondadadi et al., 2013; Zhang et al., 2016a; Yao et al., 2017a; Yan, 2022), is another instance of data-to-text generation tasks. Automated journalism systems are designed to generate news articles in real time by processing incoming data records, such as financial data or sports statistics, and moulding them into structured articles using pre-set templates.

In recent years, data-to-text generation has attracted significant research interest, with numerous large-scale datasets being proposed. These datasets, each with its unique focus or generation constraints, have introduced new dimensions to the data-to-text generation task.

One notable dataset is ToTTo (Parikh et al., 2020), a large-scale, table-to-text generation dataset constructed from Wikipedia articles. Tables in these Wikipedia articles were extracted and subsequently annotated by crowd workers, offering a distinct generation challenge. Moving from tables to biographies, the Wikibio dataset (Lebret et al., 2016) couples info-boxes extracted from Wikipedia biographies with the corresponding first sentence of each article, presenting a diverse range of entities and properties for NLG systems to navigate. The RotoWire dataset (Wiseman et al., 2017) focused on the basketball domain, comprising basketball game statistics along with their associated summaries. In the domain of biological data, the KBGen dataset (Banik et al., 2013) posits the tasks of generating text from knowledge bases, pairing sets of triples from a knowledge base with corresponding descriptive text. These triples themselves originate from biological databases that elucidate protein actions within cells. Similar to KBGen, the WebNLG dataset (Gardent et al., 2017) investigates the task of text generation from RDF triples. Lastly, the E2E dataset (Novikova et al., 2017), centred around the hospitality sector, focused on the generation of restaurant descriptions from attribute sets. The attributes encompass multiple facets of a restaurant such as name, food type, price range, customer rating, and location.

The introduction of these large-scale datasets, in their respective unique ways, has undeniably facilitated the research and application in the domain of data-to-text generation.

Existing models. Early work divides the generation process into components, each addressing a particular problem in generation. The most common breakdown of the task is a tripartite pipeline of text planning, micro sentence planning and linguistic realisation (Reiter and Dale, 1997). Some studies categorised the generation considerations into six main component problems, including content determination, document structuring, lexicalisation, aggregation, referring expression generation (REG) and surface realisation (Dale and Mellish, 1998; Reiter and Dale, 2000). These components are considered to be essential questions that have to be addressed by a complete NLG system.

In the light of recent advancements in deep learning models and the emergence of large-scale datasets, the issue of data-to-text generation has been the subject of extensive study. Despite a commonality in the deployment of neural architectures among most modern models, they display considerable divergence in terms of the overarching generation process employed. These models can be broadly divided into two categories: those following a modular pipeline approach, and those utilising an end-to-end framework.

The modular approach echoes the traditional NLG design, where the generation process is delineated into individual components, each tasked with a specific function within the workflow. Models following this approach (Gehrmann et al., 2018; Shao et al., 2019; Puduppully and Lapata, 2021; Su et al., 2021) typically involve a separate planning step, whilst utilising the neural architectures.

On the other hand, some models opt for an end-to-end architecture. Examples of these models include DataTuner (Harkous et al., 2020) and the chart-to-text model by Obeid and Hoque (2020), both of which present a neural encoder-decoder model for automatically generating natural language from structured data. Similarly, the model proposed by Marcheggiani and Perez-Beltrachini (2018) employs an encoder-decoder architecture but replaces the traditional encoder with a graph convolutional network to directly exploit the input structure.

Moreover, pre-training methods for data-to-text tasks have also been explored. Ribeiro et al. (2020) investigates the effectiveness of various pre-training strategies, while the work by Chen et al. (2020) proposes a knowledge-grounded pre-training approach, which encompasses generating knowledge-enriched text from a massive knowledge-grounded text corpus obtained from the web.

6.2 FHiG: A corpus for football highlight generation

To facilitate the exploration of the above task, I construct a football highlight generation dataset (called the FHiG dataset). Constructing a dataset like this entails addressing

several important considerations. The initial question would be: Why create a new dataset? Given the aim of this thesis to explore the feasibility of fine-grained evaluation, existing data may not suffice for the purpose of this thesis, as it may lack the desired levels of cleanliness and control. By constructing a dataset with a meticulous balance between control and real-world relevance, we can ensure precise knowledge of the data we employ and its inherent properties.

The FHiG dataset is constructed under the following requirements:

1. The data is collected in the football domain;
2. The dataset has match statistics available as the underlying world model;
3. The information (i.e., statistics) of an event is paired with a textual highlight, which can be used as the reference of the generation.

Within the football domain, data of the English Premier League (EPL) matches from season 2019/20 to 2021/22 has been curated to form the FHiG dataset. The data collection pipeline of the FHiG dataset consists of two modules: a *match statistics* collection module and a *match highlight* collection module. I will illustrate the two modules in the following two sections.

6.2.1 Collecting match statistics

The collection of match statistics includes obtaining match fixtures and retrieving match results from reliable sources, in this case the official English Premier League (EPL) website[1].

First, I obtained fixture data for each EPL season, which contains information for each football game including location, date and kick-off time, the home team, and the away team. Then I collected match results reported in the EPL results database[2], concerning events of "goal", "red card", "half time" and "full time".

6.2.2 Collecting match highlights

As mentioned in section 6.1, a *match highlight* is a sequence of words describing an important event during a football match. As the word *highlight* implies, it is usually a summary of the key information of an event. To obtain such event summaries, I collected the football tweets posted by the official English Premier League account[3] for all EPL seasons since 2019.

I consider the EPL tweets suitable for use as match highlights in this task for the following reasons:

[1]https://www.premierleague.com
[2]https://www.premierleague.com/results
[3]https://twitter.com/premierleague

- The tweets are collected from the official account of EPL. It is reasonable to believe that those tweets have a higher level of reliability and objectivity compared to reports from other unverified sources.

- In most cases, the text content of a tweet can contain up to 280 characters or Unicode glyphs. This word limit encourages the tweets reported by the EPL account to be concise and comprehensive.

- Tweets from the same account tend to be consistent in terms of vocabulary use and language style over time. This helps to make sure that our data splits will be of the same distribution.

- Most tweets from the EPL account use a diverse yet relatively formal vocabulary compared to other football-related accounts on Twitter, so they are less likely to contain undesirable features such as abusive words or colloquial phrases.

- The Twitter platform provides research-friendly APIs to retrieve historical public tweets, which facilitates the collection of useful football tweets.

For the rest of this chapter, I will use the word *highlight* to refer to a *tweet* collected from the official EPL account that describes a football event of interest.

The overall process for collecting match highlights is composed of data scraping, data filtering, tweet aligning and tweet pre-processing.

Data scraping. The Twitter research APIs facilitate easy retrieval of public tweet contents. I scraped all tweets from the official EPL account from 2019 to 2022. Apart from the main text of tweets, I also obtained relevant metadata such as twitter IDs and timestamps to facilitate later processing.

event	**tweets**
goal	John Lundstram at the double! He races in at the far post to slot in. Tammy Abraham drives a low ball through the 6-yard box and Christian Pulisic taps home. Gerard Deulofeu calmly strikes the ball down the middle to give Watford a lifeline.
red card	Christian Kabasele is shown a second yellow card for a foul on Josip Drmic. Son Heung-min is sent off for a challenge on Andre Gomes. Simon Francis is sent off after being shown a second yellow for a foul on Jota.
half time	Tammy Abraham's early lob gives Chelsea the advantage going into the break. A lively opening half with chances for both teams to find the net. Frank Lampard's men have had the bulk of possession but no way through for either side yet.
full time	Second half goals from Sergio Aguero see Man City come from behind to secure all 3 points. George Baldock cancels out Son Heung-min's opener to send the Blades up to 5th in the PL. Richarlison's volley ensures Everton pick up their first PL away win of the season.

Table 6.1: Examples of tweets describing four event types.

Data filtering. The scraped data contains all the tweets posted by the EPL account during the three EPL seasons from 2019 to 2022. I first filtered out the tweets that are

not directly relevant to the EPL matches of interest. Each match has a kick-off time and a duration, which form a *time window* during which the match happened. I extracted the *time window* for each match in the three seasons using its kick-off time and duration. Then I screened all the scraped tweets and excluded those whose timestamps do not fall into any match time window that is extracted in the above step. For the remaining tweets, I used string match to find those tweets that describe information about four event types (i.e., *goal*, *red card*, *half time* and *full time*). Figure 6.1 shows some example tweets that describe the four types of match events.

Tweet aligning. Recall that a tweet describing a match event is called a *highlight*. After the above steps, I have obtained match highlights for three EPL seasons. To find the match information that each highlight is related to, a process of aligning is performed for each highlight. The aligning process for a highlight consists of a series of steps. Firstly, the timestamp of the original tweet is used to match the highlight to a set of possible matches that kick off on the specific date. The hashtag at the end of the highlight is then used to align the highlight to a specific match (the hashtag is usually in the format of $\#AAABBB$ where AAA is the abbreviation for the *home team* and BBB is the abbreviation for the *away team*).

Tweet pre-processing. In this step, all tweets are further processed to remove URLs, emojis, etc. Unicode characters are handled by mapping a Unicode string into the closest possible representation in ASCII text. Mentions (e.g., "@premierleague") are replaced with their display names (e.g., "Premier League") on their Twitter profiles. After the pre-processing step, only the clean texts (i.e., no emojis, no URLs, no mentions) and their corresponding tweet IDs of the tweets are kept in the dataset.

To the best of our knowledge, the FHiG dataset is the first dataset that focuses on the particular task of match highlight generation in the domain of English Premier League games. The FHiG data offers the advantage of bridging the requirement for a controlled generation environment and real-world applications. This is facilitated by the semi-structured nature of its input and the availability of naturally occurring human-written text in the generation output side. The data allows for obtaining truth representations that closely approximate the real ground truth, enabling fine-grained evaluation against these representations. Consequently, the truth representations can serve as a surrogate for diagnostic evaluation, particularly in assessing faithfulness.

It is worthy noting that this match highlight collecting process can be easily applied to other EPL seasons to obtain either more training data or data from a particular year of interest.

6.2.3 Dataset statistics

Following general practice in machine learning research, I divided all the collected football data into three splits: *training*, *validation* and *testing*. The details on data statistics of the train/val/test splits are presented in table 6.2.

	FHIG$_{train}$	**FHIG**$_{val}$	**FHIG**$_{test}$
#Instances	3,011	600	600
#AvgLen	16.3	16.6	16.7
#MaxLen	39	40	35

Table 6.2: Data statistics of the FHIG dataset. Each *instance* is a (statistics, highlight) pair for a specific event. *#AvgLen* denotes the average number of tokens per highlight. *#MaxLen* denotes the maximum sentence length in each data split. The vocabulary size of the entire dataset is around 5k.

6.3 Baseline models

With the newly curated FHIG dataset, I implement baseline models of two main categories: template-based models and transformer-based models. The rationale behind the selection of these two categories is to include inherently different model architectures, in the hope that such evaluations will allow the exploration of the effectiveness of the GFD framework. This will be built on in subsequent sections.

6.3.1 Template-based models

As defined in section 6.1, given a set of information $I(x)$ about a football event x, the task of football highlight generation aims to find a generation function G that automatically generates a highlight for the event x using the information contained in $I(x)$.

The first category of generation models utilised are template-based models. Template-based generation uses pre-constructed templates and generates outputs by mapping the non-linguistic input information directly to linguistic surface forms (Reiter and Dale, 1997). I will first describe how the template libraries are built to enable template-based generation in section 6.3.1.1. With the template libraries in place, I will give details on the implementation of template-based model variants in section 6.3.1.2.

6.3.1.1 Building template libraries

I collect two libraries of football highlight templates: T_{simple} and $T_{extended}$. Each library is a list of match highlights that can be used to describe one of the four types of football events (i.e., *goal*, *red card*, *half time* and *full time*). The two template libraries differ in the number of templates they contain.

Three native speakers with good knowledge of the English Premier League were asked to write multiple sentences to describe a football event using as simple language as possible. The simple sentences are collated to form the T_{simple} template library. Then the three volunteers went on to generate more sentences for each event. At this stage, they were encouraged to use more complex vocabulary and more creative language structures to describe the events. The templates gathered at this step and the templates that are in the T_{simple} library are combined to form the $T_{extended}$ template library.

library	**event type**	**sentence length**
T_{simple}	goal	6.3
	red card	8.6
	half time	10.0
	full time (win)	9.0
	full time (draw)	10.3
$T_{extended}$	goal	8.9
	red card	11.4
	half time	15.9
	full time (win)	12.8
	full time (draw)	12.8

Table 6.3: Average sentence lengths for different types of templates collected in T_{simple} and $T_{extended}$.

After template building, the T_{simple} template library contains 6 templates for the event "goal", 7 templates for "red card", 6 templates for "half time", 6 templates for "full time (win)" when one of the teams *wins*, and 6 templates for "full time (draw)" when it is a *draw*, resulting in a number of 31 templates in total. As an extended version of T_{simple}, the $T_{extended}$ template library contains 72 templates in total, including 16 templates for "goal", 16 for "red card", 14 for "half time", and 14 and 15 for "full time (win)" and "full time (draw)" respectively. The average sentence lengths for each type of templates are summarised in table 6.3.

Table 6.4 gives examples of the templates from the simple library and the extended library used in this work. The full lists of simple and extended templates can be found in appendix A.

6.3.1.2 Generating match highlights using templates

The template-based models generate match highlights by simple slot filling in the pre-written templates. Given the event information $I(x)$, a template-based model first retrieves the *event type* e from $I(x)$. Then the model randomly samples an instance t from a set of available templates that are associated with the event type e. A target highlight y is generated by directly filling up the slots in the template t with relevant information stored in $I(x)$.

event	template
goal	*simple*: [time] - goal for [team] by [player]. *extended*: [player] guided it into the net for [team] at [time].
red card	*simple*: [time] - red card for [team]'s [player]. *extended*: [player] walked off with their head down following a red card at [time].
half time	*simple*: Half time scores: [home team] [home score] - [away score] [away team]. *extended*: The players headed into the changing room for a break, with the scores at [home team] [home score] - [away score] [away team].
full time	*simple*: Full time scores: [home team] [home score] - [away score] [away team]. [winning team] wins. *extended*: The fans of [winning team] celebrate as the final whistle is blown. [home team] [home score] - [away score] [away team].

Table 6.4: Examples of simple and extended templates for the four types of match events: *goal*, *red card*, *half time* and *full time*. Segments in square brackets (e.g., [player]) will be filled in with information from match statistics during realisation.

I calibrated two template-based models: the first model `SimTem` is built upon the simple template library T_{simple}, whilst the other model `ExtTem` uses the larger-scale template library $T_{extended}$. Experiments on the two template-based models can shed light on the effects of template size and complexity on model performance.

6.3.2 Transformer-based models

Alongside the two template-based models, I also implemented model variants based on the state-of-the-art transformer deep learning model structure (Vaswani et al., 2017). Transformer models adopt a self-attention mechanism that differently weighs the significance of each position in the input sentence. They have increasingly become the model of choice due to their high levels of performance on a wide range of NLP tasks, and their facilitation of effective training parallelisation.

Whilst template-based models utilise the original event information in $I(x)$ to directly fill up template slots, transformer-based models face a challenge in directly incorporating semi-structured statistics as input. The information within $I(x)$ needs to be **linearised** into a token sequence to enable its utilisation as input for transformer-based models.

As defined in section 6.1, the world model of an event (i.e., statistics) is a list of records that contain unique attributes and their corresponding values. The underlying statistics for an event x can be formally defined as $I(x) = \{e_1 : v_1, e_2 : v_2, \dots\}$, where $e_i : v_i$ is a record consisting of an attribute e_i (e.g., "event type") and its value v_i. The linearisation of $I(x)$ aims to translate the collection of discrete (attribute, value) pairs into a sequence representation which is suitable to use as an input to neural models. There is a large amount of literature regarding converting non-linear data into linear representations (Wiseman et al., 2017; Lebret et al., 2016; Parikh et al., 2020). I adopted a similar linearisation approach to Parikh et al. (2020). A record's internal attributes and

values are combined using a special token "<sep>". Then the record ("e_i <sep> v_i") is encapsulated in a pair of delimiters "<cell>" and "</cell>". Encapsulated records are concatenated to form a sequence. The final sequence is in the form of "<cell> e_1 <sep> v_1 </cell> <cell> e_2 <sep> v_2 </cell> ...".

On the model side, I chose two recent transformer-based models, `T5` (Raffel et al., 2020) and `Bart` (Lewis et al., 2019), and utilise multiple variants for each model for comparison. It is important to note that the models implemented in this thesis are chosen for their suitability for completing the tasks and revealing the effectiveness of the proposed evaluation framework. However, the GFD evaluation framework can be easily applied to other models as it is a model-independent framework. Details of the model variants implemented for the FHiG task are presented below:

`T5`. The `T5` model (Raffel et al., 2020) is fine-tuned on the training split of the FHiG dataset. The fine-tuned checkpoint is optimised using the Adam optimiser (Kingma and Ba, 2014) for 25 epochs with a learning rate of 1e-4 and a batch size of 16. I tried different setting for hyperparameters such as *weight decay* and the number of *warm-up steps*, and found that the values of those hyperparameters have an impact on how fast the model is trained, especially at the beginning steps, but do not have a strong impact on the final learning performance. For this reason, I set both *weight decay* and *warm-up steps* to 0 for ease of model implementation and future replication. At inference time, I used beam search (Green et al., 1977) with a beam size of 5. An early stopping strategy was adopted so that the decoding process terminates once the end-of-sentence token is generated. I also employed an n-gram similarity penalty of 3, which assures that all trigrams can only occur once in a model-generated sentence.

`T5Early`. The `T5Early` model is a `T5` model trained on the FHiG training data, but without significant hyperparameter fine-tuning or model selection. The `T5Early` model and the `T5` model are trained on the same data. The only difference between the two models is that the checkpoint of `T5` is well-selected using the validation split with extensive hyperparameter tuning, whilst `T5Early` is an intermediate checkpoint of a `T5` model that is partially trained for only 2 epochs.

`Bart`. Another baseline I implemented is based on the `Bart` model (Lewis et al., 2019). The `Bart` model is fine-tuned on the FHiG training data with extensive hyperparameter tuning and model selection. The final checkpoint stored was trained for 5 epochs with a learning rate of 1e-4 using the Adam optimiser and a batch size of 16 with an accumulation step of 4.

`BartEarly`. As the name suggests, `BartEarly` is a partly trained `Bart` model on the FHiG data without deliberate model fine-tuning. The `BartEarly` checkpoint is trained for 2 epochs with a learning rate of 1e-4.

The implementation of the four model variants is based on the HuggingFace transform-

ers library (Wolf et al., 2020). I use the *base* version of the `T5` model and the `Bart` model, initialised from their pre-trained weights. The `T5-base` checkpoint has 220 million parameters. The `Bart-base` model has 140 million parameters. The maximum input/output sequence length is set to 64. All training processes are performed on an Nvidia RTX 8000 GPU.

6.4 Evaluating grammaticality for model outputs

As previously discussed in section 3.2.3, the ERG needs to be modified to adjust to the football domain. This includes adding football-specific words and phrases to the standard ERG vocabulary, and adding mal-rules to resolve grammar conflicts with regards to subject-verb agreement rules in British English and in American English.

6.4.1 Modifications to the ERG

It is a general practice to incorporate new vocabulary when dealing with data in a new domain. I manually add common football-specific vocabulary (for example, "backpass", "own goal", "hat-trick", "stepover") to the ERG standard lexicon file to "adjust" the ERG to the specific domain of FHiG.

Another phenomenon that tends to cause failure for ERG parsing with the FHiG data is the **subject-verb agreement** in the data. This agreement error always occurs when the subject of a sentence is a *singular collective noun. Singular collective nouns* (for example, "family", "team", "committee") have the form of a singular noun, but refer to groups that are composed of individuals. Proper names of football teams (such as "Manchester City", "West Ham United") are also examples of such collective nouns.

With regard to subject-verb agreement, the rules for singular collective nouns are not always consistent. They can be treated as either singular nouns or plural nouns depending on the context, and thus are used with singular or plural forms of verbs respectively.

Generally speaking, British English and American English handle subject-verb agreement for singular collective nouns and noun phrases differently. In British English, a collective noun or noun phrase may be used as a singular or a plural noun depending on whether the speaker feels that the concept of the noun (or noun phrase) is being referred to as a single unit or as a collection of individuals, whilst in American English, a singular collective noun or noun phrase will almost always be used with a singular verb.

The standard 2020 release of the ERG treats a singular collective noun as a unit and expects a singular form of main verb to accompany it. This handling of singular collective nouns tends to trigger an agreement error when parsing the English Premier League football tweets using the ERG.

For the British dialect of English, proper name subjects such as Arsenal or Manchester United would show plural agreement with the verb. The standard version of the ERG needs to be modified to handle those plural agreement examples in FHIG. To this end, the relevant inflectional mal-rule in the 2020 version of the ERG is lifted out[4], along with the robust lexical entries for the already inflected verbs "be", "have" and "do". Such relaxation to the grammar allows more robust handling for plural agreement examples, accommodating the British setting regarding such use. The modified version of the ERG (denoted as ERG_{mod}) will be used for grammaticality evaluation for the FHIG task.

6.4.2 Parsability with the modified ERG as a proxy

Similar to the approach in section 5.4, the parsability of model outputs with the modified ERG (ERG_{mod}) is used as a proxy for grammaticality. Including the above modifications, the coverage of ERG is considerably more inclusive than without. When parsing the gold standard references (the collected tweets) using the modified ERG, I receive an overall success rate of 96.0%. I examine the sentences that do not pass the ERG parsing, and present some failed sentences and the possible reasons for parsing failure in table 6.5. The table indicates that the ERG tends to rule a sentence as ungrammatical if there is niche usage of sports specific language in the sentence (e.g., examples #1 and #2). In some cases (e.g., examples #3 and #4), the collected tweets use colloquial sentence forms which makes them not grammatical in a standard textbook sense but somewhat acceptable in a tweet reporting context.

No.	Tweet	Reason
1	Kevin De Bruyne's cross finds Raheem Sterling at the back post who heads low across goal and into the net.	missing determiner
2	Harry Kane's low ball across goal is tapped in by Lucas Moura to give the hosts the lead.	missing determiner
3	Torreira capitalises on a scramble in the box before slotting into the net.	missing object
4	Sadio Mane lashes in to give the champions an early lead.	missing object

Table 6.5: Example tweets that fail the ERG parsing. Segments that potentially caused parsing failures for the tweets are highlighted in red in the original sentences.

At first sight, the 4% gap to full coverage for the FHIG data may look like a trade-off in coverage due to the heavy reliance on hard-coded grammatical rules in the ERG. However, for any kind of automatic grammaticality evaluation, there will indubitably be false positives and false negatives. The FHIG data deals with real-world football tweets, so it is unsurprising that there are a few tweets that do not parse perfectly with the ERG.

[4]I greatly appreciate Dan Flickinger's help in providing the relaxed version of the ERG.

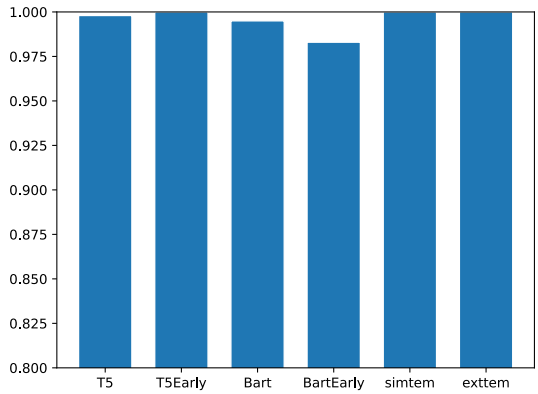


Figure 6.1: Grammaticality scores for different models.

The decisions regarding the necessity of further relaxation to the grammar rules and how much of the relaxation is viewed as appropriate need careful consideration. I argue that the important question here is not to make the evaluation module as "tolerant" as possible to any kinds of data, but to design it in a way so that the evaluation module serves as a *quantitative indicator* for the grammatical quality of an input. The ERG-based approach proposed in this work is effective enough to examine sentences for grammaticality, and it has the advantage of every sentence that passes the parsing being coupled with a list of most probable parse trees for further investigation. This makes it less of a "black-box" as all the rules that are used to make a parsing decision can be explicitly found in the ERG specifications. We know exactly what we are evaluating and the possible reasons that a sentence fails parsing.

6.4.3 Evaluation results of grammaticality

The grammaticality scores for all model variants are illustrated in figure 6.1. All models achieve a parsing success rate of higher than 0.975. This demonstrates that both the transformer-based models and the template-based models generate grammatical outputs in the majority of cases.

It is important to note that the generation models achieved higher grammaticality scores compared to the reference data, which received a grammaticality score of 0.960. One possible explanation for this disparity is that the references collected from human-written tweets inherently employ a more diverse vocabulary and incorporate a wider range of expressions when reporting on football events. As a result, they may include informal phrases, jargon, and infrequently used words, leading to a lower grammaticality score but a higher diversity measurement. This observation aligns with the diversity evaluation results presented in the later section 6.6. This suggests that a relatively lower grammaticality score does not always necessarily indicate poor performance. An extreme case of such observation is associated with the `SimTem` generation system, which achieved a perfect

grammaticality score of 1.000 but relied solely on a small set of pre-written templates for generation. This further supports the earlier assertion discussed in chapter 3 that a single metric cannot adequately evaluate all aspects of text generation quality. Instead, it is crucial to carefully examine and evaluate each generation aspect individually to gain meaningful insights about model behaviours.

It is worth noting that early checkpoints `T5Early` and `BartEarly` of two transformers models obtain grammaticality scores of 0.997 and 0.983, respectively. This result indicates that recent transformer models are able to generate grammatically well-formed language sentences from a very early stage of training.

6.5 Evaluating faithfulness for model-generated highlights

The FHiG data falls into the closed world of a restricted domain, and provides match statistics for each event as an approximate ground truth. It is therefore possible to generalise the faithfulness evaluation framework to the case of FHiG.

In this chapter, I will discuss which aspects of faithfulness will be examined for this task and explain how the intrinsic differences between the ShapeWorldICE task and the FHiG task necessitates changes in evaluation design. I will also introduce the approaches used to examine those aspects in detail and present evaluation results and analysis.

6.5.1 Another look at faithfulness

In the FHiG task, the generation of a highlight is grounded by the match statistics, which is the **world model** in this task. For faithfulness evaluation, the key information contained in a model-generated highlight is checked against the underlying match statistics. If we are to consider a model output as being *faithful* to its grounding, the information in the model output should be consistent with the underlying statistics.

What are the properties that we are interested in when evaluating the faithfulness of a football match description? If someone scores a goal, we might want to check if a model gets the name of the player correctly. We would also like to know which team the player plays for, and more essentially, if the event concerned is indeed about *goal* (not other types of events). For *full time* events, we want to check if the team names and their respective final scores are correctly generated by a model. Alongside the question of the information that we are interested in, we also need to consider if that information is verifiable using the data we have. The underlying world model (i.e., match statistics) that we have in FHiG is a **partial** ground truth. In other words, we do not have access to *all* information of the match, but a *subset* of ground truth facts demonstrated in the form of

match statistics. Therefore, the details that we investigate should be verifiable using the information contained in the world models in the FHIG dataset. Furthermore, we also need to consider how to formalise the concept of "faithfulness" so that a computational solution can be found.

Two aspects of information in this task are particularly interesting for examining faithfulness[5]: **event type** and **named entities**. I regard the most essential requirement of faithful generation to be that a model generates sentences that describe the *correct types of events*. On top of this, a generated highlight should correctly predict the named entities that contain factual information about the notable objects at an event, such as team names, player names, stadium information, etc. The first component of such a faithfulness evaluation should check the correctness of *event type* prediction in model outputs. The second component should examine the correctness of the *named entities* generated by the models.

6.5.2 A classifier for event type

The first property that I investigate in faithfulness evaluation is whether a model is able to generate match highlights that fall into correct event types. A highlight would not be considered as being *faithful* if it wrongly predicts the event type that it is supposed to describe.

To evaluate how faithful a model's outputs are in terms of event type, I implement a classifier that can predict the type of event described in a sentence. Given a highlight generated by a model, the **event type classifier** (ETC) predicts an event type label for the highlight. The event type label falls into one of the provided categories "goal", "half time", "red card" and "full time".

Consider a generation model which outputs a list of candidates $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_V$ for the test set of FHIG. For the i-th highlight $\hat{y}_i$, the ETC predicts an event type label $\hat{t}_i$ for the highlight:

$$\hat{t}_i = ETC(\hat{y}_i)$$

where $i \in 1, 2, \ldots, V$ and V is the total number of model outputs.

Here the classifier function ETC can be the `T5` model (Raffel et al., 2020), the BERT model (Devlin et al., 2019), or any other classification approaches. I implemented a pre-trained `T5` model as the backbone for the classifier due to its pioneering performance on NLP tasks. The T5-based event type classifier is fine-tuned on the FHIG training set with a learning rate of 1e-4 for 18 epochs. After fine-tuning, the event type classifier yields

[5]I also explored another aspect of faithfulness, which examines the correctness of *relations* between objects in the model outputs. However, the results from this exploration were not satisfactory. For detailed discussions on the negative results, please refer to section 6.7.2.

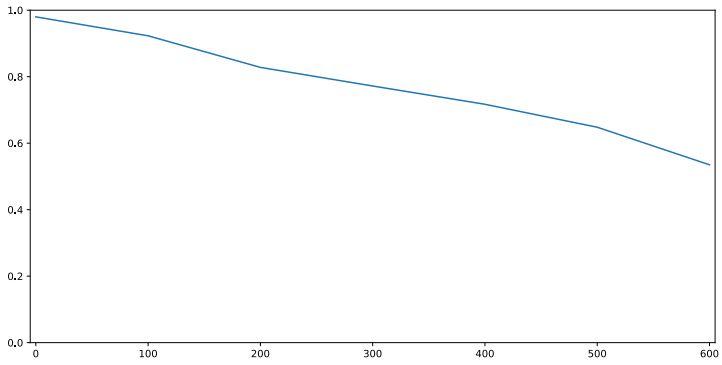


Figure 6.2: Classification accuracies predicted by the event type classifier for different shuffling choices on the reference highlights. The x-axis denotes the number of reference instances that have been randomly shuffled.

an accuracy of 0.980 on the validation data split.

To calculate the faithfulness score of a model's outputs with regards to event type, the predicted event type labels are checked against the actual event types stored in the underlying world models. The faithfulness score with regards to event type (denoted as F_{etc}) of the model outputs can then be calculated as follows:

$$F_{etc} = \frac{\#(t_i = \hat{t}_i)}{V}$$

where t_i is the actual event type expected for the i-th highlight, and V is the number of model outputs.

To verify the effectiveness of the event type classifier, I gradually shuffle the first n ($n = 0, 100, 200, \ldots, 600$) instances of the 600 reference highlights in the val split whilst keeping the remaining $600 - n$ instances in the original order. Note that when $n = 0$, all the references are kept in their original order (i.e., no shuffling). When $n = 600$, all the references are randomly shuffled.

The ETC classification accuracy of all shuffling choices is illustrated in figure 6.2. It is evident that the classification accuracy of ETC declines as the number of shuffled references increases. When all references are randomly shuffled ($n = 600$), the classification accuracy drops to 0.535, which is close to a *random guess*. This quantitative analysis justifies the credibility of the event type classifier on the FHIG-style data built in this work.

6.5.3 A hybrid model for named entity matching

The event type classifier acts as a crude sanity check for the faithfulness of model outputs, as *event type* is the most essential information that is associated with a match event. Another component of faithfulness evaluation for FHIG is a hybrid model for named entity (NE) matching between model predictions and ground truth facts.

Named entities are definite noun phrases that refer to specific types of individuals (Grishman and Sundheim, 1996; Nadeau and Sekine, 2007). The named entities of interest in the FHIG context are objects that contain information about identifiable facts in a football match. They are usually instances of named entity classes such as *person*, *location*, *organisation* and *number*.

Given a list of outputs generated by a model, the first step of faithfulness evaluation with regards to named entity matching is to extract all the named entities in those outputs. I adopt a hybrid approach which contains a pipeline of **EntityRuler**, a rule-based named entity recognition (NER) module, and **EntityRecognizer**, a statistical NER module. The implementation of both modules is based upon the pre-trained SpaCy[6] model checkpoint "en_core_web_sm". SpaCy provides an extensive NER label scheme that includes *person*, *norp* (nationalities or religious or political groups), *fac* (buildings, airports, highways, etc), *org* (organisation), *gpe* (geopolitical entities), *loc* (non-GPE locations), *product*, *event*, *work of art*, *law* (named documents made into laws), *language*, *date*, *time*, *percent* (percentage), *money*, *quantity*, *ordinal* ("first", "second", etc), and *cardinal* (numerals that do not fall under another type).

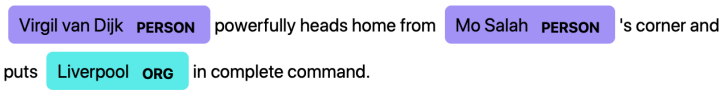


Figure 6.3: An example of a FHIG highlight and its NER results.

EntityRuler is a pipeline component in SpaCy for rule-based NER and is not trainable. To adapt the **EntityRuler** to the FHIG domain, I create a custom named entity list for football-specific entities and add them to the pattern dictionary of **EntityRuler**. The custom patterns include team names and their nicknames (e.g., "West Ham United" and its nickname "The Hammers", added as *org*), and stadium names for each team (e.g., *London Stadium* for *West Ham United*, added as *loc*). **EntityRecognizer** is a transition-based NER component that identifies non-overlapping labelled spans of tokens. Given a model-generated highlight, the rule-based NER module first scans the input text and identifies the boundaries of all textual mentions that match entries in its pre-defined NE inventory, and then identifies the types of the named entities. The statistical NER module is then run to extract named entities that missed the first round of identification by the rule-based component. Figure 6.3 gives a visualisation of the named entities identified in the sentence *Virgil van Dijk powerfully heads home from Mo Salah's corner and puts Liverpool in complete command*.

The named entities extracted from the model predictions are then compared to the match statistics (i.e., world model) of the relevant events using a relaxed *string matching*.

[6]https://spacy.io

The **named entity matching** (NEM) score (denoted as F_{nem}) is defined as the overall accuracy of named entities extracted from model outputs.

The final faithfulness evaluation reports *two* scores: F_{etc} from the event type classification module, and F_{nem} from the named entity matching module. I did not combine the two scores as the two scores investigate different aspects of faithfulness evaluation. Simply interpolating them into a single score does not make practical sense. On the contrary, as the core of this thesis is about fine-grained diagnostic evaluation of selected NLG tasks, keeping individual evaluation scores in the final reporting is useful to pinpoint specific areas of model strengths and limitations.

6.5.4 Evaluation results of faithfulness

model	F_{etc}	F_{nem}
`T5`	0.993	0.934
`T5Early`	0.958	0.874
`Bart`	0.978	0.811
`BartEarly`	0.912	0.775
`SimTem`	-	1.000
`ExtTem`	-	0.996
`REF`	0.980	0.975

Table 6.6: Faithfulness scores for different models on the FHIG data. Event type classification (F_{etc}) accuracies and named entity matching (F_{nem}) scores are reported. `REF` refers to reference highlights provided in the FHIG test set. Note that F_{etc} is only suitable for evaluating models that are trained on the same data as the FHIG references. The event type classifier requires further fine-tuning to be able to generalise to other types of text (e.g., candidates predicted by the template-based models).

The `T5` and `Bart` models are able to correctly predict most of the key information (e.g., event type, player name, team name) in their outputs. T5-based models exhibit a slightly better overall faithfulness than the Bart-based models. Early checkpoints of those two models exhibit worse results for both aspects compared to the fine-tuned models. This indicates that the faithfulness of event type and named entities increases as transformer models are trained.

A frequently observed phenomenon in T5-based and Bart-based model predictions is *hallucination*, a phenomenon that has been observed in a number of existing works (Lee et al., 2019; Rohrbach et al., 2018). Examples of hallucinated outputs from the `T5` model are listed below:

- For a *goal* event, the model often generates a phrase "*doubles [team]'s lead*". However, the numerical statement needs to be further checked and is most of the time incorrect (i.e., the new goal does not double the team's lead).

- For a *red card* event, the `T5` model sometimes generates an adverbial phrase "*after visiting the RRA*" when there is no clue in the provided data regarding if such behaviour was present before a red card was issued to a player by a referee.
- For "full time" events, the `T5` model occasionally describes the winning outcome of a team as "[team] comes from behind to claim all three points". However, a close look at the match statistics would reveal that the winning team was never "behind" the other team during the match.
- For "half time" and "full time" events, the model sometimes hallucinates about the players who scored during the match.

There is a significant difference between *hallucination* and *inference* in outputs. Both phenomena involve the generation of new information that is not directly obtained from the input. Inference is generally considered as a desirable model behaviour. Models with such ability are able to produce information that is not explicitly stated in the input, but is a logical implication of available evidences and previous knowledge. Contrarily, hallucination happens when a model generates false information based on what it picks up from the biases in the training data.

The `T5` model exhibits some level of *inference* ability. For example, when "Sheffield United" is given as the team name in the input, the model sometimes generates its abbreviated version (i.e., "Sheff Utd") in the outputs. More surprisingly, the model is able to use "the Blades" (the fan nickname for Sheffield United) as a team name to describe an event related to Sheffield United. This learnt mapping of team names and their nicknames is the desirable type of inference that we would want models to have.

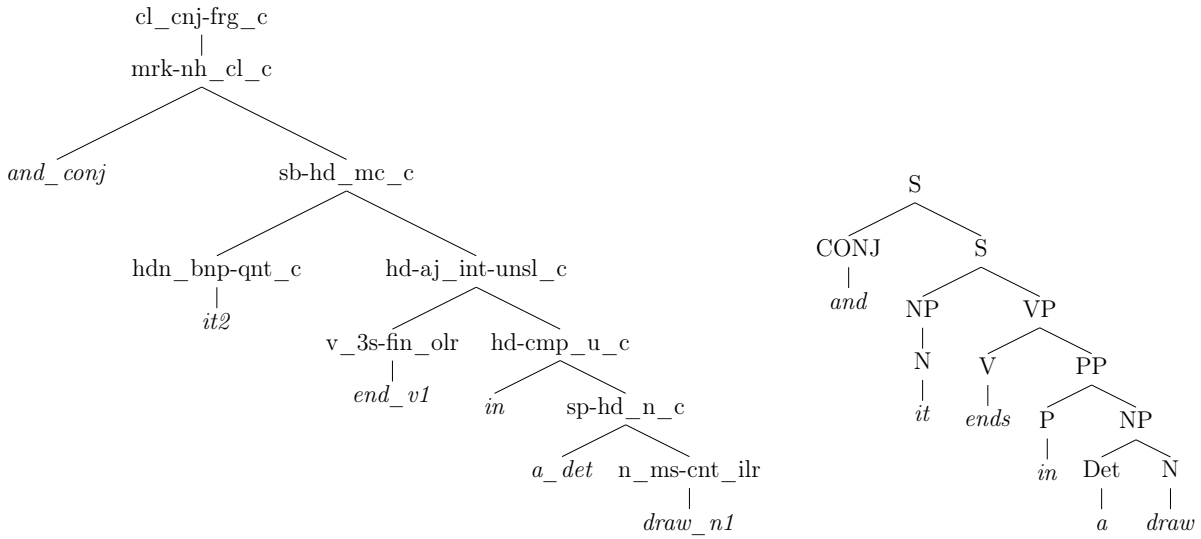
It is non-trivial to verify whether a piece of new information contained in a generated candidate is a correct inference that demonstrates desirable learning ability of a model, or a hallucinated statement where a model generates false information based on what it learnt from data bias. This is understandable, as even when humans evaluate such tasks, it is straightforward to check if some explicit information is correct, but it can become hard to decipher some out-of-scope information that might be implicitly encoded in the input.

6.6 Evaluating diversity for model-generated highlights

I focus on two levels of diversity evaluation: word-level diversity and parse-level diversity.

6.6.1 Word-level diversity

I implemented the moving average type-token ratio (Covington and McFall, 2010, MATTR) as the word-level diversity measure (refer to section 5.6.1 for details). The size of the



(a) Syntactic derivation tree.

(b) Phrase-structure tree.

Figure 6.4: Syntactic derivation and phrase structure trees for "*And it ends in a draw*".

sliding window was set to 30 when calculating the MATTR scores.

6.6.2 Parse-level diversity

Parse-level diversity can be considered as a form of syntactic diversity that operates at the construction level by comparing grammar constructions within model-generated sentences. The evaluation of parse-level diversity for FHiG follows a similar approach as for the ShapeWorldICE task, maintaining consistency in the evaluation methodology.

As described in section 5.6.2.1, model outputs are first parsed into syntactic derivation trees (Ivanova et al., 2012). These trees encapsulate important information about the grammar constructions present in model outputs. The syntactic derivation tree and its phrase structure counterpart for the sentence "*And it ends in a draw*" are illustrated in figure 6.4. *Phrasal nodes* in the syntactic derivation tree denote grammar rules exhibited in the sentence, such as *cl_ cnj-frg_ c* (fragment clause with conjunction), *mrk-nh_ cl_ c* (marker+clause), *sb-hd_ mc_ c* (head+subject, main clause), and *hd-cmp_ u_ c* (head+complement). *Leaf nodes* denote lexical entries in the input, including *and_ conj* ("and"), *it2* ("it"), *end_ v1* ("ends"), *in* ("in"), *a_ det* ("a"), and *draw_ n1* ("draw"). In the calculation of parse-level diversity, the most relevant nodes are the *phrasal nodes*. These nodes contain the entity names of specific grammar rules and indicate the locations where these rules are applied.

From the derivation trees, the grammar rules present in the model outputs are extracted. These model outputs are then mapped to vector representations, based on the presence of grammar rules in individual parse trees (refer to section 5.6.2.2 for more information). This vectorisation process converts a natural language sentence into an embedding that

represents the occurrence of grammar rules in the original sentence. Following that, the parse-level diversity, denoted as DIV_{pl}, is computed as the sparsity of the vector cluster, weighted by the number of distinct constructions observed within this output group.

6.6.3 Evaluation results of diversity

model	word-level	parse-level
T5	0.762	13.3
T5Early	0.759	14.9
Bart	0.786	11.9
BartEarly	0.755	12.7
SimTem	0.764	21.2
ExtTem	0.823	22.2
REF	0.888	29.8

Table 6.7: Diversity scores for different models on the FHiG data. Word-level and parse-level diversity scores are reported. A *window* of 30 is used to calculate the MATTR for word-level diversity. REF refers to reference highlights provided in the FHiG test set.

The word-level and parse-level diversity scores for different models are presented in table 6.7. From an initial glance, the T5 model and the early checkpoints (T5Early and BartEarly) have a word-level diversity score that is close to that of the simple SimTem model. The ExtTem model demonstrates a comparable word-level diversity to that of the human references provided in the original test data. In the collection process for the extended templates, annotators were encouraged to "think about the natural way you would describe an event". When asked this, annotators tend to use more vivid phrases to describe a specific event. For example, "hit the back of the net", "score a screamer", and "guide it into the net" are used to describe the event where a player scores. The annotators also tend to add typical information about the surrounding environment or emotional responses that can be inferred from the type of an event. For example, instead of simply saying "[player] receives a red card", annotators provide "[player] is adamant that they should not have received a red card" and "they have left [team] with a lot to do". The ability to paraphrase certain facts and add referred details accompanying events provide potential directions for future model development.

When moving onto analysing the parse-level diversity, I find that T5-based and Bart-based models yield poorer scores compared to the scores obtained from the template-based models SimTem and ExtTem. Furthermore, the fine-tuned variants for transformer-based models do not exhibit any improvement over the partly trained models, and the parse-level diversity scores fall with more training steps. A potential reason for this inverse trend in performance is a perverse incentive for the models to mimic phrases that are commonly found in the training data. As an example, there is a tendency in the T5 model outputs to

frequently repeat certain phrases such as "double someone's lead" and "give someone the lead". It has been reported in the literature (Dai et al., 2017; Lindh et al., 2018; Zhou and Lampouras, 2020) that neural models are prone to generating generic and uninteresting candidates. This is because most of these models are trained with a cross-entropy loss, which optimises the maximum likelihood of the predicted probability distribution, but also unfavourably rewards disproportionate replication of common phrases seen in the training data. A great number of approaches have been proposed to improve generation diversity (Zhou and Lampouras, 2020; Guu et al., 2018). I did not further experiment with those additions as this thesis is primarily focused on providing insights and useful tools for the *evaluation* of NLG models.

As previously emphasised, parse-level diversity calculates the sparsity of a vector cluster derived from a model's outputs, weighted by the total number of grammar constructions that the model can generate. This calculation solely relies on the model's outputs and does not require references for the test samples or specific training data for a particular task. Consequently, parse-level diversity can be compared across tasks and used as a benchmark to measure the inherent variation in generation diversity between models trained on different tasks and data. Comparing table 5.2 and table 6.7, it becomes evident that generation models trained on the real-world FHiG data exhibit a greater level of parse-level diversity in their outputs compared to models trained on the ShapeWorldICE data, which was generated using a fixed set of pre-determined construction rules.

6.7 Negative results and failed attempts

In this section I report some failed attempts and negative results on FHiG. Although the content in this section does not directly contribute to the final thesis, I decided to include these discussions here as they might be useful for people who are working on similar data-to-text tasks or across related domains.

6.7.1 Data augmentation attempts

When curating the FHiG data, I tried multiple methods to increase the dataset size. A straightforward approach is to scrape more EPL tweets from previous seasons, as discussed in section 6.2. Another method that I tried is **back-translation** (Federmann et al., 2019; Sennrich et al., 2016; Edunov et al., 2018). Back-translation is a widely used **paraphrasing** technique that can be used for data augmentation. It works by translating instances in a language into an intermediate language, then translating the translations back into the original language. It is an effective method for generating similar instances for existing data (Suzuki et al., 2017; Thompson and Post, 2020; Zhou and Bhat, 2021). I choose *English-French-English* and *English-Chinese-English* translation pipelines due to

the fact that machine translation between those language pairs generally yield a better performance benefiting from large amounts of parallel training corpora.

sent 1	It is a nightmare start for Norwich as Grant Hanley inadvertently diverts Divock Origi's cross into his own net.
en-fr	***f***: C'est un début de cauchemar pour Norwich puisque Grant Hanley détourne par inadvertance la croix de Divock Origi dans son propre filet. ***b***: It's a nightmare start for Norwich since Grant Hanley inadvertently hijacks Divock Origi's cross in his own net.
en-zh	***f***: 这是Norwich的噩梦开始，因为Grant Hanley无意中将Divock Origi的十字架到自己的网里。 ***b***: This is the beginning of Norwich's nightmare because Grant Hanley accidentally moved Divock Origi's cross into his network.
sent 2	Raheem Sterling produces a clinical low finish after Kevin de Bruyne breaks forward in the middle and puts him through.
en-fr	***f***: Raheem Sterling produit une finale clinique basse après Kevin de Bruyne en avant au milieu et le met à travers. ***b***: Raheem Sterling produces a low clinical final after Kevin de Bruyne in the middle and puts it through.
en-zh	***f***: Raheem Sterling在Kevin de Bruyne从中间冲过去并让他通过之后，产生了一个低临床完成量的临床完成量。 ***b***: Raheem Sterling, after Kevin de Bruyne ran through the middle and let him pass, produced a low clinical completion.

Table 6.8: Examples of back-translation for two FHiG sentences. **en-fr** denotes English-French translation, and **en-zh** denotes English-Chinese translation. ***f*** refers to *forward* translation, i.e., from English to the other language. ***b*** refers to *back* translation, i.e., from the other language back to English.

I implement back-translation models using the transformer "translation" pipeline (Wolf et al., 2020), with model checkpoints specifically trained for English-French and English-Chinese translation tasks[7] (Tiedemann and Thottingal, 2020).

However, the outcome of back-translation is far from satisfactory. Table 6.8 illustrates a few examples from the translated football highlights. Football-specific vocabulary is a case where back-translation tends to fail. Most general-purpose translation systems are trained on general corpora of web articles that contain articles of all domains. However, some words that are commonly seen in football reports (such as "foul", "strike", "fire", "far corner", "finish", "feed") are polysemic and have specific meanings in the context of football, which can be different from their uses in non-football texts. For example, "strike" as a noun in the football context means "the action of kicking a football, especially hard so that it travels a long distance". Another common use of the word refers to "the action

[7]The model checkpoints used for English-French back-translation are *Helsinki-NLP/opus-mt-en-fr* and *Helsinki-NLP/opus-mt-fr-en*. The model checkpoints used for English-Chinese back-translation are *Helsinki-NLP/opus-mt-en-zh* and *Helsinki-NLP/opus-mt-zh-en*.

to refuse to continue working because of an argument with an employer about working conditions etc"[8]. For this polysemic ambiguity, a general-purpose neural translation model may fail to understand the context of the sentence "*What a strike!*", as the word "strike" can be interpreted either way when the domain constraint is not sufficiently encoded in the translation process.

To address this problem, translation models that can better accommodate football language are needed. Although there is little existing work in football-specific translation models, there has been some attempts at pre-training sports-specific language models. SportsBERT (Srinivasan and Mashetty, 2020) is a BERT-based transformer model trained on sports news articles, covering a wide range of domains such as basketball, hockey, football, cricket, soccer, tennis, etc. Its tokeniser is trained to include more sports-related tokens to its vocabulary. A potentially fruitful method to explore would be to use SportsBERT's pre-trained weights to initialise current neural translation models, and then fine-tune translation models on FHiG-style parallel data. As this exploration requires extensive work in constructing large amounts of multilingual parallel data in the football domain, I will mark this exploration as future work.

6.7.2 An information extraction (IE) based method for relation triple matching

In the faithfulness evaluation in section 6.5, I first look at the correctness of *event type*, and then evaluate *named entities* (such as *team*, *player*). I wanted to explore another aspect of faithfulness, which examines the correctness of *relations* between objects described in the model outputs. More specifically, I am interested in examining the **subject-predicate-object** (SPO) triples enclosed in the model outputs. To this end, I use the state-of-the-art OpenIE toolkit (Angeli et al., 2015) to extract SPO triples from model outputs, and evaluate the triples against the underlying world models by relaxed triple matching.

Note that the representation used to encode objects and relations is not hugely important here. It can be in the form of SPO triples, parse trees or abstract meaning representations (Banarescu et al., 2013, AMRs). I choose to extract SPO triples using OpenIE because the SPO triples predicted by OpenIE directly use the words in the original sentences (without over-abstraction), which allows certain level of surface information to be kept in the triples. This was expected to facilitate direct fact-checking of objects and their relations.

However, this attempt was not satisfactory. The main bottleneck lies in OpenIE's insufficient extraction of SPO triples. For long sentences with clauses, OpenIE often ignores part of the sentences, which causes significant information loss from the missing segments. For example, for sentence *A sensational opening half for Liverpool, who are*

[8]Definitions obtained from the Cambridge Dictionary. https://dictionary.cambridge.org.

out of the blocks for the 2019/20 PL season in some style, OpenIE only extracts a single triple (*block*, *be in*, *style*) from the dependent clause, completely ignoring the information contained in the first half of the sentence. This indicates that the existing information extraction tools are not good enough for the FHiG data curated in this work. This is potentially due to the lack of training data in the particular football domain for existing tools. I will leave the investigation of better information extraction methods to future work, as this is beyond the scope of this thesis.

6.8 Discussion

In this chapter, I explored the application of the GFD evaluation framework within the closed domain of football highlight generation. Unlike the fully synthetic setup of ShapeWorldICE, the FHiG dataset aimed to test the generalisation ability of the GFD evaluation framework in real-world scenarios.

The evaluation of grammaticality exemplified the challenges posed by the inherent variability of language use and its nuanced interpretations. While a grammaticality evaluation using the ERG was achievable for the FHiG data, significant modifications were necessary to handle football-specific vocabulary and the divergent handling of subject-verb agreement in British English and American English.

The empirical findings from the faithfulness evaluation indicated that despite having a partial ground truth representation (i.e., world models), evaluating faithfulness remained a challenging task and required substantial adjustments to the evaluation approach.

Diversity evaluation showed significant variation in lexical and syntactic diversity among outputs generated by different models. However, none of the model outputs were comparable to the diversity exhibited in human-written references. Transformer-based models achieved lower diversity scores compared to template-based models, likely due to the generic generation commonly observed in neural generation models trained with maximum likelihood estimation (MLE) objectives (Dai et al., 2017; van Miltenburg et al., 2018).

The empirical results also highlighted the subtle tension between grammaticality and diversity. Whilst human-written references exhibited a higher level of diversity than model-generated sentences, they received the lowest grammaticality score. A possible explanation for this disparity is that the references collected from human-written tweets inherently employ a more diverse vocabulary and incorporate a wider range of expressions when reporting on football events. As a result, they may include informal phrases, jargon, and infrequently used words, leading to lower grammaticality scores but higher diversity measurements. Similar results were also observed in the GFD evaluation for the subsequent TIAGE task (see chapter 7).

Furthermore, the contrast between the parse-level diversity measurement outcomes for the ShapeWorldICE task and the FHiG task underscores the impact of training data on the diversity of model outputs. Specifically, models trained on the real-world FHiG data demonstrated a higher level of diversity (at both the word level and the parse level) in their generated text. In contrast, models trained on the ShapeWorldICE data, which was synthesised using a fixed set of construction rules, exhibited lower diversity levels.

It is worth noting that while this chapter primarily focused on the football highlight scenario, the methodology employed here is not restricted to this particular use case. It can be readily extended and applied to other closed-world scenarios, provided there is a reasonable level of access to underlying world models.

Chapter 7

TIAGE: Topic-shift dialogue generation and evaluation

This chapter expands on the use of the GFD protocol for NLG evaluation to another scenario: open-domain chit-chat dialogue generation with a specific focus on topic shifting[1]. In the following, I will first introduce the task of topic-shift dialogue generation in section 7.1, followed by a thorough discussion of related work in section 7.2 and an introduction of the TIAGE dataset specifically curated for this task (see section 7.3). The baseline models utilised in this task are introduced in section 7.4. Finally, the GFD evaluation approaches implemented for this particular task are described in sections 7.5-7.7.

7.1 The task of topic-shift dialogue generation

Dialogue generation in NLP refers to the task of generating human-like responses in a conversational setting. It involves creating a system that can understand the context of a conversation and generate appropriate and coherent responses. This field has witnessed significant progress over the years, driven by advancements in natural language processing,

[1]Acknowledgement: The work presented in this chapter builds upon my conference paper titled "TIAGE: A Benchmark for Topic-Shift Aware Dialog Modeling", which was accepted and published in the Findings of EMNLP 2021, co-authored with Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu and Ann Copestake (Xie et al., 2021).
I would like to acknowledge the contributions of my co-authors in shaping this research. Chenyan Xiong played a significant role by suggesting the research question and engaging in valuable discussions about potential exploration directions. Zhenghao Liu provided assistance with model parameter tuning inquiries. Zhiyuan Liu and Ann Copestake contributed valuable comments and participated in discussions that influenced the final narrative of the paper. All experiments and analyses presented in the paper represent my own work.

The published paper primarily focused on introducing topic-shift modelling in dialogue settings. In this chapter, I further expand upon the paper by addressing the following aspects: 1. The adaptation of the GFD evaluation protocol to the topic-shift dialogue generation setting; 2. Providing a comprehensive account of evaluation approaches and presenting empirical results; 3. Conducting an extensive survey of existing dialogue datasets, which serves as an additional contribution of this chapter.

Figure 7.1: An example of topic-shift behaviours in human conversations. Topic-shift utterances are highlighted in green and in *italic*. Changing the topic helps keep the conversation going on.

machine learning, and deep learning techniques.

Dialogue generation tasks can be broadly divided into two categories depending on the level of domain constraint: *task-oriented* dialogue which focuses on assisting users with specific tasks, such as making restaurant reservations or booking flights, and *open-domain* chit-chat which involves creating conversational agents that can engage in open-ended conversations with users. The topic-shift dialogue generation task in this thesis falls under the broader umbrella of *open-domain* dialogue generation. The generation of a response in this task is still open-ended and is domain independent, and is in the form of chit-chat. In this case, the specific task constraint is a specific emphasis on topic shifting - the generated response should not only continue the dialogue, but also be able to shift the dialogue topic in another direction. This particular requirement differentiates this task from the general setting of dialogue generation, and represents a novel angle of inquiry in the context of GFD evaluation.

The task of topic-shift dialogue generation can be formally defined as generating a topic-shift response s_{TS} given a dialogue history X_T. In this formal format, given a specific turn T in a dialogue, we denote all its previous utterances as the *context* $X_T = x_1, \ldots, x_i, \ldots, x_N$ where x_i is the i-th utterance in the dialogue context, and N is the context length. The task of *dialogue generation* can be considered as automatically generating a *response* given the context X_T. Denote a *topic-shift response* as $s_{TS} = s_1, \ldots, s_i, \ldots, s_L$ where s_i is the i-th token in the response and L is the sentence length. Similarly, an *on-topic response* is denoted as $s_{NTS} = \bar{s}_1, \ldots, \bar{s}_i, \ldots, \bar{s}_M$ where $\bar{s}_i$ is the i-th token in the response and M is the sentence length. As stated earlier, the emphasis of this dialogue generation task is to generate a response that changes the current conversational topic, whilst maintaining the overall flow of dialogue. This particular requirement represents a specific constraint on open-domain dialogue generation, testing a model's ability to swiftly and naturally change

topics in a dialogue.

This task presents a significantly different challenge to the previously discussed tasks. Firstly, chit-chat dialogue generation involves weaker grounding from the input content, as conversation participants have more freedom to steer the conversation in a multitude of directions. In comparison, the input information in the other tasks provides a stronger grounding for generated content. Secondly, this dialogue task specifically focuses on modelling topic-shift behaviour in a dialogue setting, which imposes a specific task requirement, greatly influencing the generation process. These factors influence the criteria for determining high-quality generated outputs, providing fruitful results for investigation in the context of GFD evaluation.

7.2 Related work

This section first provides an overview of existing literature on *topic* and *topic shift*. Then, I survey existing dialogue datasets, presenting their basic statistics and features.

7.2.1 Topic and topic shift

7.2.1.1 Seeking a working definition of "topic"

In human conversations, Brown and Yule (1983) observed that speakers tend to "speak topically" by aligning their contributions with recent elements of the discussion. However, the existence of an agreed-upon, working definition of "topic" remains uncertain.

The definition of "topic" has been a subject of extensive debate in linguistics. Early discussions of this notion rooted in psychological and cognitive terms. von der Gabelentz (1869) referred to "topic" as the "psychological subject" in a conversational interaction. He equated the psychological subject of a sentence with what a speaker desires the listener to think about: "und ich nenne das, woran, worüber ich den Angeregten denken lassen will, das *psychologische subjekt*" (von der Gabelentz, 1869). Some linguists described *topic* as "the centre of shared attention". According to Gundel et al. (1993), the entities in focus at a given point in discourse are likely to continue as topics in subsequent utterances. Similarly, Chafe (1994) defined topic as the totality of semi-active information at a particular time. Another perspective, proposed by Berthoud and Mondada (1995), argued that discourse topic can be defined in terms of "aboutness", representing what a portion of the interaction is about.

In the realm of communication, the concept of "topic" typically refers to the entity that a speaker identifies as the subject of the given information in conversation. Within this context, *topic* and *comment* are usually considered as interconnected concepts, where the *topic* represents the subject or theme which is being discussed, and the *comment*

represents the information which is being expressed about that topic. This interconnected relationship is commonly referred to as a *topic-comment structure* in the literature.

Despite extensive debates attempting to define the concept of topic, a recurring observation found in the literature is the prevalent use of the term without a definition that goes beyond a lay or intuitive understanding. Goutsos (1997) asserted that there is a significant lack of consensus on the notion of topic. Grobet (2002) shared this criticism directed towards the literature studying topic, highlighting its reliance on weak, vague and intuitive definitions.

It is important to note that in this thesis, instead of attempting to provide a precise definition or formalisation of the concept of "topic" in conversations, the focus is on exploring a specific aspect that is inherent in human conversations: *topic shift*. The subsequent section will delve into the discussion of related work on topic shift.

7.2.1.2 Topic shifts in human conversations

Human conversations naturally encompass multiple topics and exhibit fluent transitions between them. In real-life scenarios, people change topics in conversations for various reasons. They may change topics to avoid lengthy discussions, dodge personal questions, or when a current topic triggers discomfort. In a chatbot scenario, the need to shift topics usually arises when the chatbot has exhausted the available information on a particular topic or has a specific goal to accomplish, such as booking a flight or providing recommendations to the user.

Wardhaugh (1985) argued that conversations typically involve shifts from one topic to another, sometimes even mixing multiple topics simultaneously. This implies that speakers in a conversation do not restrict themselves to a single topic throughout; instead, they tend to shift to new or related sub-topics. Brown and Yule (1983) described topic shift as the change of subject within a conversation, where speakers introduce new topics to enhance the interest and effectiveness of the discussion. Similarly, McCarthy (1991) emphasised the significance of topic shifting in sustaining conversation flow and avoiding silence.

Identifying topic changes can be challenging as it often relies on the analyst's intuition. Extensive research has been dedicated to investigating strategies and markers that indicate topic shifts. Early studies in discourse analysis (Brown and Yule, 1983) suggested that a variety of adverbial phrases, such as "what about", "on a different note" and "changing the subject" are frequently used as *discourse markers* to indicate topic changes in conversations. These markers can also be used as features for a discriminative classifier (Galley et al., 2003) or observed variables in a probabilistic model (Dowman et al., 2008). However, it is important to note that in practice, the discourse markers that are most indicative of topic change often vary significantly depending on the specific domain of the data (Purver, 2011). This dependency on domain poses a limitation for methods that solely rely on

Dataset	Year	Category	Data source	Annotations	#Dialogues	#Turns per dialogue
CALLHOME American English Transcripts (Kingsbury et al., 1997)	1997	telephone conversation	telephone conversation transcripts	speaker details, channel quality, etc	120	N/A
Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000)	2000	audio transcript	audio recordings and transcripts	metadata about speaker details	60	N/A
ISL Meeting Transcripts (Burger et al., 2002)	2002	group meeting	meeting transcripts	annotations for spontaneous speech events and disfluencies	112	N/A
ICSI Meeting Corpus (Janin et al., 2003)	2003	group meeting	meeting recordings and transcripts	dialogue acts, speech quality	91	N/A
SLX Corpus of Classic Sociolinguistic Interviews (Strassel et al., 2003)	2003	interview	audio recordings transcripts	sociolinguistic variable survey	8	N/A
NIST Meeting Pilot Corpus (Garofolo et al., 2004)	2004	group meeting	meeting transcripts	metadata about topics, meeting forums, participants, recording conditions, etc	19	N/A
NPS Internet Chatroom Conversations (Forsyth and Martell, 2007)	2007	chit-chat	online chatrooms	dialogue acts, PoS tags	15	N/A

NXT Switchboard Annotations (Calhoun et al., 2009)	2009	telephone conversation	telephone conversation transcripts	focus/contrast, dialogue acts, prosodic annotations	2,400	N/A
MPC Corpus (Shaikh et al., 2010)	2010	chit-chat	crowd-souring	communication links, dialogue acts, local topics, meso-topics	14	520
Twitter Conversation Dataset (Ritter et al., 2010)	2010	micro-blog conversation	Twitter	N/A	1,300,000	N/A
Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011)	2011	movie script	movie scripts	characters, metadata about genre, release year, cast lists, etc	220,579	N/A
Ubuntu Dialogue Corpus (Lowe et al., 2015)	2015	task-oriented	Ubuntu chat logs	N/A	930,000	8
Reddit Corpus (Schrading et al., 2015)	2015	micro-blog conversation	Reddit	subreddit categories	21,133	N/A
OpenSubtitles (Lison and Tiedemann, 2016)	2016	movie script	movie/TV show subtitles	metadata about release year, language, duration, genre, etc	322,000	N/A
DailyDialog (Li et al., 2017)	2017	chit-chat	English learners' writings	dialogue acts, emotion classes	13,118	7.9
Frames (El Asri et al., 2017)	2017	task-oriented	crowd-sourcing	dialogue acts	1,369	14.6

DSTC7 (Galley et al., 2019)	2018	micro-blog conversation	Reddit	N/A	2,364,228	<2
Persona-Chat (ConvAI2) (Zhang et al., 2018a)	2018	chit-chat	crowd-sourcing	personas	10,907	14.8
MultiWOZ (Budzianowski et al., 2018)	2018	task-oriented	crowd-sourcing	dialogue states, dialogue acts	8,438	13.68
Empathetic Dialogues (Rashkin et al., 2019)	2019	chit-chat	crowd-sourcing	emotion labels	25,000	4.31
Wizard of Wikipedia (Dinan et al., 2019)	2019	chit-chat	crowd-sourcing	topics	22,311	9
Topical-Chat Dataset (Gopalakrishnan et al., 2019)	2019	chit-chat	crowd-sourcing	annotations about reading set utilisation and sentiment	11,319	22
DialoGPT (Zhang et al., 2020b)	2019	micro-blog conversation	Reddit	N/A	147,116,725	<2
Blended Skill Talk (Smith et al., 2020)	2020	chit-chat	crowd-sourcing	skill annotations regarding knowledge, empathy, personal situations and personal background	∼5,000	11.2

Table 7.1: Survey of existing dialogue datasets. *#dialogues* denotes the number of dialogues contained in a dataset, and *#turns per dialogue* denotes the average number of dialogue turns in a dialogue in each dataset.

these markers, as it can be challenging to adapt them to new domains or settings.

7.2.2 Survey of existing dialogue datasets

I conducted an extensive survey of dialogue datasets released before the year 2021, and summarised their features and statistics in table 7.1. Please note that the surveyed datasets exclusively focus on English dialogue datasets. Although there are undoubtedly valuable dialogue datasets curated for other languages, this thesis specifically concentrates on language generation tasks using the English language. It is important to clarify that this survey does not aim to provide a complete overview of all dialogue datasets. Instead, its purpose is to serve as a resource for dialogue researchers who are seeking suitable data to support their research work in the field of dialogue generation.

Early dialogue-related datasets primarily consist of transcripts from spontaneous phone conversations (Kingsbury et al., 1997; Calhoun et al., 2009) and meeting recordings (Burger et al., 2002; Janin et al., 2003; Garofolo et al., 2004). These transcripts are usually automatically transcribed from audio recordings and then checked by human annotators. However, they often contain incomplete utterances resulting from interruptions during phone or meeting conversations. Many of these datasets provide annotations that can be used for discourse analysis, such as dialogue acts, named entities, etc.

Other sources of dialogue data include subtitles and scripts from movies or TV shows (Lison and Tiedemann, 2016; Danescu-Niculescu-Mizil and Lee, 2011), and online conversation threads from micro-blogging service platforms (Schrading et al., 2015; Ritter et al., 2010; Zhang et al., 2020b). With the advent of crowd-sourcing techniques, numerous large-scale conversational datasets have been curated, each with its respective specific task focus. In domain-specific and task-oriented dialogue settings, conversations revolve around a predefined domain or goal (Lowe et al., 2015; Budzianowski et al., 2018). In such scenarios, the chatbot is designed to operate within the specified domain or objective, ensuring it stays aligned with the predetermined requirements. On the contrary, other datasets (Forsyth and Martell, 2007; Shaikh et al., 2010; Li et al., 2017; Zhang et al., 2018a) concentrate on open-domain dialogue, typically in a chit-chat environment. These works differentiate themselves by their unique research focuses, such as incorporating external knowledge-base grounding (Gopalakrishnan et al., 2019), incorporating persona features (Zhang et al., 2018a), or combining multiple dialogue generation skills (Smith et al., 2020).

The surveyed works comprise a wide range of conversation forms. Of these, a significant number of the datasets listed in figure 7.1 comprise multi-party conversations, which consist of dialogues among a group of individuals engaged in discussions on specific issues. These conversations can take place in various settings, including in-person meetings (Burger et al., 2002; Janin et al., 2003) or online chatrooms (Forsyth and Martell, 2007; Shaikh et al., 2010). Other datasets consist of two-party conversations, which can be in the form of

Persona 1	Persona 2
I am an engineering student	I am a mother
I enjoy spending time with my friends	I have a son and a daughter
I really enjoy watching classic films	I love the Harry Potter books
My mother is the greatest chef in the world	Winter holidays have always been my favourite
I stand out among my surroundings	I enjoy the outdoors

Table 7.2: Example persona profiles selected from the Persona-Chat dataset.

phone conversation transcripts (Kingsbury et al., 1997; Calhoun et al., 2009), social media conversation threads (Ritter et al., 2010; Schrading et al., 2015), or specially crowd-sourced dialogues (Zhang et al., 2018a; Rashkin et al., 2019; Dinan et al., 2019; Smith et al., 2020). These conversations typically involve interactions between two individuals and serve as valuable resources for studying various aspects of dialogues characterised by a bipartite structure.

7.3 The TIAGE dataset

In this section I introduce the TIAGE dataset. I start off by explaining why I consider the Persona-Chat dataset (Zhang et al., 2018a) as a suitable data source for this project. Topic-shift labels were manually annotated based on the Persona-Chat dataset to facilitate the topic-shift dialogue task. I will also discuss the human annotation process of the topic-shift labels in TIAGE, and analyse data statistics and conversational patterns in this dataset.

7.3.1 The Persona-Chat dataset

Introduction to Persona-Chat. The Persona-Chat dataset (Zhang et al., 2018a) is a chit-chat dialogue dataset with pre-scripted persona profiles. It was introduced as part of the Second Conversational Intelligence Challenge (ConvAI2). The dataset aims to enhance dialogue modelling by incorporating user personas. Each dialogue in the dataset consists of two participants, each assigned a *persona profile* - a set of sentences describing a user's background, interests and preferences. The dataset comprises 1,115 unique pre-defined personas, ensuring results comprise a wide range of user characteristics and preferences. Examples of persona profiles used in Persona-Chat are presented in table 7.2. During data collection, annotators were required to carry out conversations based on the persona profiles they were assigned. A dialogue consists of multiple conversational turns between the users, with each turn connecting two utterances from separate participants. Dialogues in the Persona-Chat dataset cover a large variety of topics and have a minimum dialogue length of 6 turns. These characteristics combine to make this dataset suitable for analysing topic-shift behaviours in conversational dialogues.

The entire Persona-Chat dataset contains a total of 162,064 utterances over 10,907 dialogues, 131,438 utterances (8,939 dialogues) of which are used for training, 15,602 utterances (1000 dialogues) for validation, and 15,024 utterances (968 dialogues) for testing.

Persona-Chat has been instrumental in developing personalised dialogue systems and improving the contextuality of conversational agents (Wolf et al., 2019; Liu et al., 2020). It has been widely used to train and evaluate dialogue systems, providing a benchmark assessment of dialogue models' ability to generate coherent and contextually appropriate responses, taking into account the user's persona and goals.

Rationale for the choice of data source. I constructed TIAGE by augmenting the Persona-Chat dataset (Zhang et al., 2018a) with topic-shift human annotations. I consider Persona-Chat as a suitable dataset for topic-shift annotation for the following reasons:

1. The Persona-Chat data was collected online in a textual form by mimicking chit-chat scenarios. As such, it reflects real-world conversations and the chit-chat setting allows for more topic shifting turns in dialogues, which aligns with the focus of this research.

2. Dialogues in this dataset contain more than 10 dialogue turns, and longer dialogue contexts tend to exhibit a conversational flow with more topics.

3. Despite the fact that some participants in Persona-Chat may have rushed into changing topics to quickly exchange their profile information, I observed that most of the participants still manage to change topics in a natural and coherent way, making this dataset a sensible choice to study topic-shift behaviours.

While acknowledging that Persona-Chat is a suitable choice for this project, it is important to recognise that it is not without limitations. For example, the dataset contains typos, and informal language usage such as slang, abbreviations, and misuse of punctuation. These aspects have an impact on the evaluation of grammaticality (discussed in section 7.5).

7.3.2 Annotation process

To annotate the topic shifts within the data, I randomly selected 500 dialogues (7,861 dialogue turns) from the original Persona-Chat dev/test datasets. I asked annotators to examine each dialogue turn and indicate whether the conversational topic is changed in that turn. In the annotation pool, there were a total of 25 human annotators who are fluent in English and have sufficient linguistics knowledge on topic shift behaviours in discourse studies. Each dialogue was randomly distributed to, and annotated by two independent annotators.

	Dialogue 1	TS Label
[Speaker1]	Hi! How are you this evening?	N/A
[Speaker2]	Good. I spent all afternoon walking my dogs. I've three Labradors.	**0**
[Speaker1]	Cool, that's a lot of dogs. Do you like music? I love it.	**1**

	Dialogue 2	TS Label
[Speaker1]	I think you are great. You are my best friend.	N/A
[Speaker2]	My best friend is a bear, bears don't have friends, that's why they're my favourite.	**0**
[Speaker1]	Webster's dictionary defines weddings as the fusing of two metals with a hot torch.	**1**

Table 7.3: Annotated dialogue examples in TIAGE.

During the annotation process, all annotators were given identical annotation guidelines, and were talked through the general aim of this annotation task (without injecting too much unnecessary information to avoid unfavourable bias). They were then asked to annotate each dialogue turn indicating whether they think the conversational topic has changed. One consideration for the annotation process was the fact that topic is co-constructed. As such it is very limiting to analyse a turn for itself when trying to identify topic transitions. To facilitate the recognition of slowly transited topics, annotators were encouraged to take into account both the previous two turns and the following two turns of the target dialogue turn when making a decision. This assisted decision making for cases where topics are slowly developed and transited. Appendix B gives more details of the annotation guidelines used in the annotation process for TIAGE.

Following the annotation process, I obtained a dialogue dataset with gold standard topic-shift labels at each dialogue turn. The Cohen's Kappa score (Cohen, 1960) for all annotations is 0.48[2]. Table 7.3 showcases examples of labelled dialogues selected from TIAGE. To the best of our knowledge, TIAGE is the first dataset that focuses on topic-shift behaviours in open-domain dialogue data.

7.3.3 Dataset analysis

This section presents dataset statistics of TIAGE and an analysis of the topic-shift patterns observed in the annotated TIAGE dialogues.

[2]The Cohen's Kappa score falls within the range of 0.41 to 0.60, indicating a moderate level of agreement, which confirms the quality of human annotations in TIAGE.

	WeakSupo$_{train}$	WeakSupo$_{dev}$
#Dialogues	7,939	1,000
#Instances	108,711	13,788
#AvgTurns	14.7	14.8

(a) The weak supervision data split.

	Anno$_{train}$	Anno$_{dev}$	Anno$_{test}$
#Dialogues	300	100	100
#Instances	4,767	1,546	1,548
#AvgTurns	15.6	15.5	15.6

(b) The human-annotated data split.

Table 7.4: Data statistics. *#AvgTurns* denotes the average number of turns per dialogue. Each *instance* is a (context, response) pair around a specific dialogue turn. The average number of tokens per utterance is 11.8. In the human-annotated data split, the average number of topic-shift turns per dialogue is 3.5. The vocabulary size of the entire dataset is around 18K.

7.3.3.1 Dataset statistics

As shown in table 7.4, TIAGE provides weak supervision data and human-annotated data to train dialogue models. Weak supervision data is selected from the original Persona-Chat training set and helps adapt NLG models to Persona-Chat-style data. The weak supervision data consists of 8,939 dialogues and is split into two sets: WeakSupo$_{train}$ and WeakSupo$_{dev}$. Human-annotated data consists of 500 annotated dialogues with topic-shift annotations at each dialogue turn. I split them into 300 Anno$_{train}$, 100 Anno$_{dev}$ and 100 Anno$_{test}$ dialogues respectively. As each dialogue has multiple dialogue turns, I extracted (context, response) pairs as instances for all turns in each dialogue.

7.3.3.2 Analysis of topic-shift patterns

From examining the range of topic-shift utterances labelled by human annotators in TIAGE, I observed an interesting pattern of [comment; topic shift]. In this pattern, the response that initiates the change in topic typically begins with a brief comment on the previous dialogue context, followed by a sentence that introduces a different conversational focus. The comment usually corresponds to the sentiment previously expressed in the dialogue. This pattern resonates with findings in pragmatics research (Brown and Levinson, 1987; Goldsmith, 2007). Correspondingly, according to Leech (1983), there exists a politeness principle that aligns with conversational maxims formulated by Paul Grice (Grice, 1975). Leech outlined six specific maxims: tact, generosity, approbation, modesty, agreement, and sympathy. Politeness, particularly tact, involves strategically avoiding conflicts and demonstrating consideration for others.

Among the six maxims, two are particularly relevant to the observed topic-shift patterns

in TIAGE. The approbation maxim emphasises expressing beliefs that convey approval of others, seeking to make others feel good by demonstrating solidarity. The other maxim, the sympathy maxim, aims to maximise sympathy and minimise antipathy between oneself and others. This involves employing speech acts such as expressing commiseration and offering condolences. These actions also align with Brown and Levinson (1987)'s positive politeness strategy, which emphasises attending to the interests, wants, and needs of the listener.

Where these concepts can bring value to the discussion of topic-shift patterns in TIAGE is in the formulation of new topic responses. When speakers introduce a new topic, it is a common positive politeness strategy to first respond to the content uttered by other speakers (Leech, 1983, 2014). This pattern holds potential value for dialogue systems seeking to generate topic-shift utterances in a natural and coherent manner. Prior to introducing a new topic, it is favourable for dialogue systems to first generate a comment regarding the previous topic that expresses either approbation (e.g., "great", "that's cool") or sympathy (e.g., "that's too bad" or "I'm sorry to hear that"). This demonstrates the systems' attentiveness to the users' interests and needs.

7.4 Baselines

I built a `T5-NLG`$_{TS}$ response generator using the pre-trained T5 model (Raffel et al., 2020). I initially trained the T5 model on the WeakSupo$_{train}$ data for transfer learning. After this step, the model is more accustomed to the specific Persona-Chat language style. I then further fine-tuned the model on the topic-shift instances (i.e., where topic shifts occur) in the Anno$_{train}$ data for specific topic-shift adaption.

I simultaneously ran this task on a number of topic-insensitive NLG models to provide a basis for comparison. These baselines were chosen due to their state-of-the-art performance on dialogue tasks. I trained a `T5-NLG` model on the WeakSupo$_{train}$ data without any topic-shift signals. I also implemented the `DialoGPT` model (Zhang et al., 2020b) fine-tuned on the same data as another baseline.

For implementation, I used the small version of DialoGPT and the base version of T5. My implementation was based on the HuggingFace Transformers library (Wolf et al., 2020). All models were optimised using the Adam optimiser (Kingma and Ba, 2014) with a learning rate of 5e-5 and a batch size of 64. I set the maximum input sequence length to 512. The training was carried out using 1 Nvidia RTX 8000 GPU and took around 15 hours.

No.	Reference response	Reason
1	So are you enjoy fall?	verb tense
2	I like that show to. Grey's Anatomy is my favorite.	mixed up spelling
3	I afraid of snakes, you?	sentence fragment
4	I love to go running, it's my favorite passion.	comma splice

Table 7.5: Examples of reference responses that failed the ERG parsing. Segments that potentially caused parsing failures for utterances are highlighted in red in the original sentences. A *sentence fragment* is an incomplete sentence used in place of a complete sentence, typically lacking either the subject or predicate necessary to make it an independent clause. *Comma splice* refers to an inappropriate language use in English where a comma is used to join two clauses that have no appropriate conjunction.

Models	G	$\mathbf{F}_{ts}$	$\mathbf{DIV}_{wl}$	$\mathbf{DIV}_{pl}$
`DialoGPT`	0.962	0.457 (0.509)	0.787	18.7
`T5-NLG`	0.984	0.527 (0.456)	0.677	13.3
`T5-NLG`$_{TS}$	0.987	0.540 (0.544)	0.697	14.2
Ref	0.895	0.632 (1.000)	0.826	23.2

Table 7.6: GFD evaluation scores obtained by baselines for the topic-shift dialogue generation task. G denotes the grammaticality score. F_{ts} denotes the faithfulness score obtained using the `RetroTS-T5` classifier. Scores in parentheses represent faithfulness results obtained through human evaluation regarding topic-shift occurrences. DIV_{wl} denotes the word-level diversity score, whilst DIV_{pl} denotes the parse-level diversity score. Note that evaluation scores for reference responses (in the row **Ref**) were calculated only for comparison purposes. The GFD framework remains a reference-less approach, meaning that the evaluation for models with the GFD does not rely on any references.

7.5 Evaluating grammaticality of model-generated responses

As mentioned earlier, the Persona-Chat dataset, which the TIAGE dataset is built upon, was collected through crowd-sourcing in a casual chit-chat setting. Consequently, the language usage in this dataset reflects the way people communicate in informal online conversations. This distinction sets this corpus apart from the data I examined in the previous two tasks, namely image captions generated using a formal semantic engine and football tweets from the official EPL account.

7.5.1 Parsability with the ERG as a proxy

Continuing in line with the previous two tasks, the grammaticality of model-generated dialogue responses is evaluated by assessing the parsability of these sentences with the ERG.

I parsed the reference responses in the TIAGE test set, and obtained a success rate of

89.5%. This indicates that 89.5% of the references were successfully parsed using the ERG. Table 7.5 illustrates a few examples of reference responses that failed the ERG parsing. Note that these references were directly chosen from the original Persona-Chat dataset and were created by crowd workers in the data collection process of Persona-Chat. The reference instances that failed the parsing commonly exhibited grammatical mistakes such as:

- Verb tense and form. Some ill-formed utterances exhibit errors in verb tenses or forms. For example, the sentences "so are you enjoy fall?" (example #1 in table 7.5) and "do you exercising a lot?" demonstrate mixed-up tenses and incorrect verb forms, resulting in ungrammatical sentences.

- Sentence segment. A *sentence segment* refers to an incomplete sentence that is used in place of a complete sentence. Typically, an incomplete sentence lacks a subject or a main verb, which is necessary for it to function as an independent clause. In some cases, a sentence segment may have a subject and a main verb, but other elements of the sentence are missing, making the sentence grammatically incorrect. Example #3 in the illustration table is an example of this common grammar mistake.

- Run-on sentence or comma splice. A run-on sentence refers to a sentence that continues without proper punctuation or conjunctions to separate two independent clauses. A comma splice refers to the incorrect usage of a comma to join two clauses that lack an appropriate conjunction. Example #4 in the table is an example of comma splice mistakes. This particular error is the most commonly observed in the Persona-Chat utterances.

- Typo. Another type of error commonly found in the reference sentences is typos. For instance, confusion between words such as "to" and "too", as well as "an" and "and", can lead to ill-formed sentences.

As previously discussed, the Persona-Chat dataset was collected through a crowd-sourcing platform, where annotators followed given instructions to complete annotation tasks. Due to the informal nature of the online chit-chat setting, and the inherent data quality issue associated with crowd-sourced datasets (Aker et al., 2012; Daniel et al., 2018), the sentences in the resulting Persona-Chat dataset exhibit various imperfections. These include spelling errors, incorrectly used punctuation, incomplete sentences, and other grammar-related mistakes. These empirical findings highlight that the use of the ERG to measure grammaticality is not limited to assessing model outputs, but can also provide an insight into the grammar-related quality of human-written "gold-standard" references.

7.5.2 Evaluation results of grammaticality

The second column in table 7.6 presents the grammaticality scores obtained by baseline models. It is evident that all three models - `DialoGPT`, `T5-NLG`, and `T5-NLG`$_{TS}$ - achieved grammaticality scores of more than 96.0%. This indicates that the baseline models are able to generate grammatically well-formed sentences for the TIAGE generation task. The T5-based models (`T5-NLG` and `T5-NLG`$_{TS}$) exhibit slightly higher grammaticality scores than the DialoGPT model. Fine-tuning the `T5-NLG`$_{TS}$ model specifically on topic-shift utterances resulted in only a slight effect on the grammaticality of the generated responses, as indicated by the grammaticality measurement results of 0.984 and 0.987 for the `T5-NLG` and `T5-NLG`$_{TS}$ models, respectively.

It is worth noting that the grammaticality score of the reference responses written by human annotators is lower than those generated by the baseline models. This difference can potentially be attributed to the imperfect data quality inherent to crowd-sourced data, as discussed in the previous section.

7.6 Evaluating the faithfulness of model-generated responses

The faithfulness evaluation for the previous two tasks, image captioning and data-to-text generation, primarily focuses on *grounding faithfulness*. That is, the input information of the two tasks imposes strong groundings for the generation process, and the faithfulness evaluation of a generated sentence is evaluated by examining its consistency with the grounding information. However, in the context of a general chit-chat dialogue, dialogue history only poses weak grounding for what should be generated as a response. Most existing dialogue research examines this weak grounding of dialogue history through generation qualities such as relevance and coherence. Using these metrics to qualify faithfulness aims to reveal how well a generated utterance is connected with the overall context in the dialogue history, and they are valuable for measuring on-topic dialogue generation.

However, in the specific case of topic-shift dialogue generation, a model is required to generate an utterance that deliberately changes the current conversational topic. This posits a specific challenge, in that an ideal response should have a different focus to the dialogue history, making the evaluation of relevance and coherence less relevant in this particular case.

To this end, the faithfulness evaluation proposed for the TIAGE task focuses on *task faithfulness*. More specifically, as the task requirement is to generate a *topic-shift* response, the faithfulness evaluation examines whether a model-generated utterance fulfils

the essential requirement of the task, i.e., generating a response that changes the current dialogue topic. In other words, a model-generated sentence is considered as faithful to the task if the sentence does change the conversation topic to a different direction.

7.6.1 A classifier for topic shifts

For the dialogue contexts $X_1, X_2, \ldots, X_V$ contained in the test set of ANNO_{test}, a generation model outputs a list of candidates $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_V$ for the V dialogue contexts, respectively. The topic-shift classifier is designed to automatically predict whether there is topic shifting occurring at a specific dialogue turn. Formally, for the i-th generated response $\hat{y}_i$, the topic-shift classifier (denoted as TSC) predicts a label $\hat{t}_i$ based on X_i and $\hat{y}_i$:

$$\hat{t}_i = TSC(X_i, \hat{y}_i) = \begin{cases} 1, \text{if there is a topic shift between } X_i \text{ and } \hat{y}_i \\ 0, \text{otherwise} \end{cases} \tag{7.1}$$

The faithfulness score (denoted as F_{ts}) of a dialogue model's outputs with regards to topic-shift occurrence can be defined as:

$$F_{ts} = \frac{\sum_{i=1}^{V} \hat{t}_i}{V}$$

where $\hat{t}_i$ is the topic-shift label predicted by the topic-shift classifier for the i-th pair of dialogue context and response, and V is the number of model outputs.

I implemented three baseline models as the topic-shift classifier. The first model, `GenEnc` uses the GEN encoder (Zhang et al., 2019a) to separately encode dialogue context and response into embeddings to estimate the topic-shift intents. `GenEnc` uses a cosine similarity threshold of 0.25 to filter out (context, response) pairs, and classify them as topic-shift occurrences. The other two classifiers were developed utilising the state-of-the-art BERT and T5 models. Specifically, I implemented a `BERT-wiki727` model (Devlin et al., 2019) trained on the WIKI-727K dataset (Koshorek et al., 2018). I also employed a T5 model (Raffel et al., 2020) fine-tuned on the ANNO_{train} data with topic-shift labels as the retrospective T5 topic-shift classifier (denoted as `RetroTS-T5`). I used the base version of BERT and T5 models, initialised from their pre-trained weights.

Topic-shift classifiers were first tested on the human-annotated ANNO_{test} split. The classification performance of the three classifiers is presented in table 7.7. Precision, recall, and F1 scores were also computed for human-generated annotations for comparison. The human annotations attained an F1 score of 0.644, illustrating the inherent complexity in judging topic-shift occurrences in conversations. Of the three classifiers, `RetroTS-T5` significantly outperformed the others and matched the level of human performance. This suggests that topic shifts in Persona-Chat dialogues exhibit certain patterns that the `RetroTS-T5` classifier can capture from the topic-shift annotations. Consequently, I will

Model	Precision	Recall	F1 score
GenEnc	0.337	0.199	0.250
BERT-wiki727	0.412	0.020	0.038
RetroTS-T5	0.631	0.673	0.651
Human	0.687	0.607	0.644

Table 7.7: Classification performance obtained by the topic-shift classifiers. The evaluation results associated with **Human** in the bottom row were reported under the situation where annotations from one annotator were used as gold standard references and were compared with annotations from the other annotator.

use the RetroTS-T5 classifier for the automated evaluation of faithfulness in this task.

Given that the highest-performing classifier for detecting topic-shift occurrences, RetroTS-T5, achieved only an F1 score of 0.651 on the ANNO$_{test}$ split, its use as a calibrator for faithfulness evaluation may potentially introduce errors into the evaluation process. With this in mind, I carried out an additional human evaluation to assess topic-shift occurrences, with the aim to examine and validate the effectiveness of the topic-shift classifier in evaluating faithfulness for the TIAGE task.

7.6.2 Evaluation results of faithfulness

The faithfulness scores achieved by the three NLG models - DialoGPT, T5-NLG and T5-NLG$_{TS}$ - are presented in the third column of table 7.6. Scores in parentheses correspond to human evaluations for topic-shift occurrences. Results from human evaluation indicate that only about half of the responses generated by the three models exhibit a shift in topic, with T5-NLG$_{TS}$ producing the most topic-shift utterances. This observation is echoed in the evaluation results derived from the RetroTS-T5 classifier. However, the classifier failed to correctly identify topic-shift occurrences in the **Ref** responses. These reference responses, consisting entirely of utterances that initiate a change in topic, should logically have a perfect faithfulness score of 1. This, unfortunately, was not captured by the RetroTS-T5 classifier. There is a notable gap between the faithfulness score predicted by RetroTS-T5 and the result from human evaluation for the reference responses, suggesting a deficiency in the reliability and effectiveness of the RetroTS-T5 classifier.

7.7 Evaluating diversity of model-generated responses

When undertaking the diversity evaluation for this task, the most appropriate approach was on the same lines as the diversity evaluations for the previous two tasks. As such, I investigated the syntactic diversity of model-generated dialogue responses through two vectors: word-level diversity and parse-level diversity.

7.7.1 Word-level diversity

I consider word-level diversity as the first level of diversity analysis. Word-level diversity, denoted as DIV_{wl}, is defined as the normalised type-token ratio (TTR) of model-generated responses (refer to equation 5.2 in section 5.6.1). The size of the sliding window was set to 30 when calculating the MATTR scores.

7.7.2 Parse-level diversity

The parse-level diversity, denoted as DIV_{pl}, examines the diversity of syntactic constructions in model-generated sentences. It is calculated in the same way as in equations 5.4-5.6 in section 5.6.2. A higher value of DIV_{pl} indicates a greater parse-level diversity in the outputs of a model, indicating a broader range of grammatical constructions.

7.7.3 Evaluation results of diversity

The word-level and parse-level diversity scores obtained by the baseline models are presented in the last two columns of table 7.6. Both the word-level and parse-level diversity scores of the `DialoGPT` model were higher than the two T5-based models. The word-level diversity score for the human-written reference sentences was 0.826, which surpassed the word-level diversity scores of all the baseline models. Moreover, the references achieved a parse-level diversity score of 23.2, which was over 60% greater than the score obtained by the $\texttt{T5-NLG}_{TS}$ model. These findings suggest that the dialogue responses written by humans generally exhibit a wider range of word usage and a more diverse set of grammatical constructions in comparison to the outputs produced by the GPT-based and T5-based generation models.

7.8 Other tasks with TIAGE

In addition to the topic-shift dialogue generation task, I further investigated another task: *topic-aware dialogue generation*. Although this work is not directly related to the GFD evaluation of this thesis, I decided to include the explorations here as they illustrate the potentials of using the TIAGE data for other topic-shift related dialogue scenarios.

7.8.1 Topic-aware dialogue generation

Given a dialogue context X_T, the goal of the topic-aware dialogue generation task is to generate a topic-shift response s_{TS} if a change of topic is considered as favourable, or an on-topic response s_{NTS} if otherwise. The topic-aware dialogue generation task differs from the topic-shift generation task in that topic-aware generation asks models to identify the need to change topics by themselves, and then generate topic-shift or on-topic

Model	Precision	Recall	F1 score
TSManager	0.340	0.170	0.220
RetroTS-T5	0.709	0.657	0.682
Human	0.687	0.607	0.644

Table 7.8: Classification performance obtained by the TSManager model, compared to that of the RetroTS-T5 classifier. The evaluation results associated with **Human** in the bottom row were reported under the situation where annotations from one annotator were used as gold standard references and were compared with annotations from the other annotator.

Model	BLEU-2	METEOR	ROUGE_L	CIDEr
DialoGPT	0.063	0.077	0.134	0.125
T5-NLG	0.082	0.087	0.159	0.175
TADial	0.082	0.087	0.162	0.177

Table 7.9: Evaluation results of topic-aware dialogue generation on all instances in ANNO_{test}.

responses according to the generated prediction, whilst topic-shift generation aims to generate topic-shift responses given any dialogue context.

Models. Three baselines were built for comparison. The first two baselines are T5-NLG and DialoGPT models fine-tuned on the WEAKSUPO_{train} data.

Alongside the construction of these models, I implemented a pipeline system (denoted as TADial) with the particular purpose of generating topic-aware responses. The first step is to decide whether a change of topic is necessary at a specific dialogue turn. To predict topic-shift signals based on the dialogue context X_T, I implemented a T5-based topic-shift manager (denoted as TSManager) and fine-tuned it on the ANNO_{train} data. The major difference between TSManager and the RetroTS-T5 classifier introduced earlier is that RetroTS-T5 has access to both the dialogue context and the response, whilst TSManager makes topic-shift predictions based solely on the context.

Subsequently, I went on to train two T5-based response generators: T5-NLG_{TS} and T5-NLG_{NTS}. The two response generators produce either a topic-shift or an on-topic response by switching between one another, guided by the topic-shift signals from TSManager. T5-NLG_{TS} aims to generate topic-shift responses, and T5-NLG_{NTS} is fine-tuned on non-topic-shift instances to generate on-topic responses.

Results. I first examined the TSManager classifier by its performance when predicting topic-shift labels. Table 7.8 reports the classification results of the TSManager classifier on the ANNO_{test} set. Looking at the results, there is a clear gap in classification performance between RetroTS-T5 and TSManager. I posit the reason for this being that the predictive setting of TSManager is inherently more difficult than RetroTS-T5, as it is asked to predict topic-shift labels based solely on dialogue context.

Following these results, I then tested TADial and two topic-insensitive baselines on all

instances in ANNO_{test}. The results from table 7.9 indicate that `TADial` with a dedicated topic-shift management component does not yield better performance than the `T5-NLG` model, which is simply trained on dialogue instances with no topic-shift labels. This indicates that, mostly likely due to the deficiency of `TSManager` signals, hard-wiring a topic-shift management component into a generation pipeline falls short at the task of improving generation results. It remains a challenging task to produce well-timed and good-quality topic-shift signals based on dialogue context only, which hinders overall topic-aware dialogue generation.

It is worth noting that the evaluation of the topic-aware generation task is based on n-gram methods, including BLEU-2, METEOR, ROUGE_L and CIDEr. I did not further explore GFD evaluation for this task, as the primary purpose of these experiments was to demonstrate the versatility of the TIAGE annotations across various scenarios. The intention was to provide preliminary explorations that could shed light on potential use cases with TIAGE. A more in-depth examination of model evaluation will be left for future work.

7.9 Discussion

This chapter explores the application of the GFD framework to the TIAGE benchmark, which focuses on topic-shift response generation in an open-domain chit-chat setting. This setting posed a challenging environment for the GFD evaluation framework, particularly for the faithfulness evaluation. Whilst the evaluation of grammaticality and diversity could be effectively generalised to the case of TIAGE, the adaption of the faithfulness assessment was less satisfactory.

The evaluation of grammaticality of the generated responses indicated that machine learning models can produce grammatically well-formed sentences with high consistency. An interesting revelation was the relatively lower grammaticality score achieved by the human-written references. This could be attributed to the colloquial language typically used in casual dialogues, complicated by the inherent data quality issues often found in crowd-sourced datasets (Hossain and Kauranen, 2015; Daniel et al., 2018).

This chapter also examined the faithfulness of model-generated responses, which took a slightly different approach compared to the previous tasks. Instead of focusing on *grounding* faithfulness, the TIAGE task necessitated an evaluation of *task* faithfulness, considering the topic-shift constraint imposed by the task. I operationalised the faithfulness evaluation of topic-shift generation through analysing topic-shift occurrences, thus translating the task requirement of generating topic-shift responses into a measurable parameter. Despite these efforts, the faithfulness evaluation did not reach a satisfactory level of reliability. The empirical findings presented in section 7.6 indicated the substantial challenges and

limitations in applying the faithfulness evaluation in the GFD framework to the particular task of TIAGE.

The diversity evaluation revealed a significant contrast between human-written and machine-generated responses. Human-written responses demonstrated wider vocabulary use and more diverse syntactic constructions. This observation aligns with the findings from the FHIG task.

Chapter 8

Conclusion

8.1 Summary of main ideas and results

This thesis explores the evaluation of NLG tasks, proposing a decomposition approach to evaluating NLG models for grammaticality, faithfulness, and diversity (GFD). The versatility of the GFD framework is empirically investigated using three distinct NLG tasks: synthetic image captioning, football highlight generation, and topic-shift dialogue generation.

In chapter 1, I raised several pivotal questions that this thesis aims to explore. In concluding this thesis, I will provide some reflections on these questions, drawing on the investigations undertaken in the preceding chapters.

Evaluation in the context of NLG. Evaluation, at its core, involves the systematic application of scientific criteria to ascertain the value and merit of a subject. It serves two primary purposes: it affords insights into the subject being evaluated, and it facilitates reflection on existing methodologies, thus directing future research paths. This thesis delves into the intricacies of NLG evaluation, and suggests a dissection of the evaluation process into individual contributors. The decomposition approach employed in this thesis operates along two orthogonal axes:

Firstly, the focus of the evaluation is broken down into three distinct qualities: grammaticality, faithfulness, and diversity (GFD). Each of these qualities represents essential aspects of text generation outputs. Among the three evaluation aspects, grammaticality and diversity focus on lexical and syntactic properties of a text, whereas faithfulness examines the alignment of a model-generated text with the intended content and task requirements.

Secondly, to investigate the adaptability of the GFD framework across different NLG scenarios, I applied the framework to three distinct tasks: synthetic image captioning (ShapeWorldICE), football highlight generation (FHiG), and topic-shift dialogue gen-

eration (TIAGE). The three NLG tasks encompass vastly different domains, leading to inherent variations in language usage. Furthermore, the grounding level of task inputs and the availability to the underlying ground truth differ significantly across tasks, presenting varying degrees of challenge for evaluation. Specifically, the synthetic configuration of ShapeWorldICE provides a fully controlled environment for GFD evaluation. The FHiG task serves as a logical next step from the synthetic image captioning task, as it involves real-world inputs whilst providing partial access to the underlying ground truth. The TIAGE task represents a specific form of dialogue generation, with an emphasis on generating topic-shift responses. This task presents an intriguing and challenging scenario for GFD evaluation due to its open-domain chit-chat nature and the imposition of a topic-shift constraint in task requirements.

By decomposing NLG evaluation into examining three text qualities on three distinct tasks, a grid structure of evaluation is formed, accounting for the variations in generation requirements and task differences. This grid organisation enables a detailed, systematic examination of generation qualities across various NLG scenarios.

The choice of essential qualities within an evaluation framework. In evaluation, different stakeholders may hold divergent views on what constitutes the "merit" of a subject. Seen from this perspective, the selection of evaluation aspects is inherently a contested matter. However, there are still fundamental guidelines that should be followed when conducting NLG evaluation. Specifically, the selection of granular qualities should be guided by the core objective of facilitating a meaningful evaluation.

As a starting point towards addressing this problem, I attempted to present an overview of notable factors that could potentially contribute to overall text quality (refer to section 2.1.1). Among these factors, I highlighted three specific aspects - grammaticality, faithfulness and diversity - that serve as fundamental calibrators to gauge the quality of model-generated text. Furthermore, I have operationalised these three concepts such that computational approaches can be devised to assess how well these aspects are represented in the text.

The main contribution of this thesis is the principled GFD framework with an operationalised evaluation of grammaticality, faithfulness, and diversity. This is the outcome of my attempt to discover commonality within evaluation across various generation cases, whilst appreciating the inherent nuances within NLG evaluation. In my perspective, breaking down NLG evaluation into investigating these three qualities represents a meaningful initial effort towards a "glass-box" NLG evaluation.

Generalisation ability of the proposed framework. One of the key questions that I sought to answer throughout this research was the feasibility of a comprehensive, automated evaluation of the three salient, well-defined metrics across carefully selected tasks, and

to what extent these metrics can be generalised across such tasks. The grammaticality and diversity evaluations demonstrated a commendable level of generalisation across tasks, albeit with necessary adaptions, such as modifications to the ERG which was utilised for the grammaticality and diversity assessments. However, the faithfulness evaluation presented a significant challenge. Despite considerable efforts in tailoring it to fit individual tasks, the empirical results demonstrated in chapters 5-7 suggested that establishing a universal evaluation schema for assessing faithfulness remains an extraordinarily challenging task, even with carefully delineated task designs and focused evaluation objectives.

As the generation environment aligns more closely with real-world scenarios, it becomes increasingly difficult to obtain complete access to the underlying ground truth representations. Consequently, the task of evaluating generation systems becomes significantly more challenging. This is particularly the case for the faithfulness evaluation. In a well-controlled environment such as the ShapeWorldICE task, evaluating the faithfulness of a model's outputs to a grounding world model is straightforward. However, when it comes to text generation from semi-structured data, the partial availability of the underlying truth representations leads to a partially achievable assessment of faithfulness. Furthermore, in an open-domain chit-chat setting, the task of defining faithfulness becomes extremely complex in itself. Even with the intentional task constraint of topic-shifting and an attempt to translate the evaluation of faithfulness into the analysis of topic-shift occurrence as a means of operationalisation, the task remains non-trivial due to the lack of underlying truth models. This required resorting to manual annotation as a form of evaluation, despite which the faithfulness evaluation was still far from satisfactory and failed to provide reliable insights regarding the faithfulness in model outputs. This underscores the inherent complexity of NLG evaluation and the need for extreme caution when addressing the problems it presents. It calls for meticulous examination and thoughtful consideration in both defining and operationalising evaluation metrics, particularly when the grounding truth becomes less accessible or when working in more complex real-world scenarios.

Another dimension of generalisation concerns the applicability of the GFD framework across languages. Although initially developed for English, the GFD evaluation framework proposed in this thesis is inherently language-independent and can readily be adapted for other languages, provided that suitable grammar engines are available.

Insights gained from the evaluation attempts. An empirical analysis of the three tasks and their corresponding generation models indicates that state-of-the-art neural models possess the capacity to generate grammatical outputs after undergoing training for a small number of epochs on the training data. The disparity in parse-level diversity scores between the synthetic ShapeWorldICE data and the real-world datasets (FHiG and TIAGE) suggests that the linguistic diversity inherent in the training data significantly influences the constructional diversity of a model's outputs. In contrast to the

real-world datasets, ShapeWorldICE captions, generated from a fixed set of pre-defined construction rules, exhibit limited language variability, leading to a decrease in the diversity of outputs generated by models trained on such data.

The faithfulness evaluation arguably poses the greatest challenge when applying the GFD evaluation framework to real-world generation tasks. While adjustments can easily be made to the evaluation of grammaticality and diversity to suit a specific task, the assessment of faithfulness is not easily generalisable across tasks, especially in situations where the access to the representation of the underlying ground truth is restricted.

Additionally, as empirically demonstrated in section 5.5, current holistic automatic metrics (such as BLEU and SPICE) present certain limitations in assessing the faithfulness of model outputs. The effectiveness of BLEU as an NLG metric has been questioned in existing literature (Reiter, 2018), particularly for tasks where the output content is not narrowly constrained, like dialogue (Liu et al., 2016). However, this thesis does not suggest the GFD evaluation framework as a replacement for existing automatic evaluation methods. Instead, it underscores the importance of reflecting on the prevailing NLG evaluation paradigm and offers an initial attempt to delineate the contributors to text quality and their evaluation across unique generation environments. The work presented in this thesis should be seen as a complementary evaluation framework to standard practices. When combined with diagnostic NLG datasets, the GFD framework can provide valuable insights into model capabilities and limitations, enhancing standard evaluations.

This thesis also highlighted the subtle tension between grammaticality and diversity. For instance, a higher grammaticality score achieved by a generation model compared to human-written references does not necessarily indicate superior generation outputs, but rather points to a potential lack of diversity in the generated sentences. It is therefore crucial to recognise that relying solely on a specific evaluation metric may not provide a comprehensive understanding of the actual model performance.

Furthermore, this thesis empirically demonstrated the inherent differences between rule-based evaluation approaches and statistical evaluation approaches, a topic addressed in multiple sections of previous chapters (see section 3.2.2 and section 7.6). Rule-based evaluation approaches, such as the ERG parsing for grammaticality evaluation, use pre-defined rules and are principally data-independent. However, the adaption of rule-based metrics to a new domain requires expert knowledge, which is non-trivial to incorporate in many scenarios. In contrast, statistical evaluation approaches, such as the event type classifier for FHiG, can remedy these issues, as they can be automatically trained using specifically curated data. However, the reliability of evaluation methods in this genre is greatly affected by the quality and the representativeness of the data that is used to train the evaluation models. Although the construction of these statistical models does not necessarily rely on domain-specific expert knowledge, it does require large amounts of

suitable data to be curated. There is no one-size-fits-all answer regarding which category of evaluation methods is most useful. The selection of evaluation methods depends on the target domain and the availability of pre-defined rules or high-quality training data.

This thesis also emphasises that evaluation design should be considered at every phase of a research project, starting from its inception. The choice of evaluation metrics warrants thorough examination and discussion during the experimental stage and should be carefully explained and analysed in any subsequent publications. This level of transparency is vital to ensure replication across different studies. It is through the pursuit of more refined definitions and detailed dimensions of NLG evaluation that we can establish a foundation for meaningful model comparisons. Without such commonality, our comparisons would be akin to comparing apples with oranges, lacking consistency and congruity.

8.2 Looking forward

The work presented in this thesis provides a provisional agenda towards a fine-grained, systematically structured NLG evaluation. This section outlines several prospects for expanding upon the work detailed in the previous sections.

To begin with, it is worthwhile to consider alternative approaches for implementing the three metrics. For instance, in the GFD evaluation framework, the ERG served as the underlying grammar engine for text parsing, which was subsequently employed in metric assessment. In future research, an insightful avenue for exploration would involve experimenting with different dependency grammars, such as the Universal Dependencies (Nivre et al., 2020, UD v2), to enhance our understanding of the GFD framework's robustness. Another promising area for investigation pertains to the parse-level diversity evaluation. In this thesis, the calculation of parse-level diversity scores solely relied on the *occurrence* of language constructions found in parse trees. A valuable extension could involve incorporating factors like *frequency* and the *hierarchy* of parse tree nodes in the diversity assessment, offering a more refined perspective on diversity measurement.

The scope of the GFD evaluation framework can be further expanded. The current research examined three qualities, grammaticality, faithfulness, and diversity, which are essential requirements for good-quality generation outputs. However, the evaluation framework can be enhanced by incorporating additional qualities, such as relevance and informativeness (see discussion in section 2). Incorporating these concepts into the existing GFD framework would be a highly desirable addition.

Next, the phenomenon of hallucination exhibited in neural model outputs could be further investigated. Differentiating between hallucinations and plausible inferences, as explored in sections 3.3.3 and 6.5.4, remains a compelling research challenge, deserving of more intensive scrutiny. The possibility of penalising for hallucination in neural model

outputs with the faithfulness metric could also be further explored.

Furthermore, the evaluation framework outlined in this research remains relevant in the context of large language models (LLMs). The GFD framework can be readily applied to evaluate synthetic texts generated by LLMs. In addition, the diversity evaluation approach introduced in this study focuses on analysing word usage and text constructions. This approach has the potential to effectively differentiate between texts generated by distinct groups, such as LLMs, native first-language writers (L1) and second-language learners (L2). The topic of identifying human-generated vs model-generated texts has been actively explored in recent workshops and shared tasks[1].

Moreover, the GFD evaluation results offered insights into model limitations, providing a useful compass for subsequent model inspection and development. For instance, in the ShapeWorldICE generation task, I reported the evaluation results of two existing models on three basic visual compositions. Future efforts could delve into discerning at what point image captioning models start to exhibit learning ability on each type of ShapeWorldICE dataset, and the influence of strategies such as data blending and curriculum learning on model performance, as investigated for the VQA task in Kuhnle et al. (2018). Furthermore, empirical findings from the TIAGE experiments uncovered that it remains a challenging task for dialogue models to predict high-quality topic-shift signals based solely on dialogue context (see section 7.8.1). Assessing the necessity and appropriateness of initiating a topic change is inherently a challenging task, thus presenting an appealing research question to probe further. It is my aspiration that my work on the TIAGE task, besides providing a platform for testing the GFD evaluation protocol - which is a contribution to this thesis in itself, also lays the groundwork for further exploration of topic-shift tactics in dialogue generation models in the NLG field.

[1]For instance, the ALTA 2023 Shared Task: https://www.alta.asn.au/events/sharedtask2023/. The goal of this task is to build automatic detection systems that can discriminate between human-authored and synthetic text generated by LLMs.

Bibliography

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1955–1960, 2016.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4971–4980, 2018.

Jean Aitchison. Language change. In *The Routledge Companion to Semiotics and Linguistics*, pages 111–120. Routledge, 2005.

Ahmet Aker, Mahmoud El-Haj, M-Dyaa Albakour, Udo Kruschwitz, et al. Assessing crowdsourcing quality through objective tasks. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1456–1461, 2012.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7558–7567, 2019.

Yara Alharahseheh, Rasha Obeidat, Mahmoud Al-Ayoub, and Maram Gharaibeh. A survey on textual entailment: Benchmarks, approaches and applications. In *Proceedings of the 13th International Conference on Information and Communication Systems*, pages 328–336, 2022.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 382–398, 2016.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 936–945, 2017.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.

Larry Andrews. *Language Exploration And Awareness: A Resource Book For Teachers*. Routledge, 2013.

Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5561–5570, 2018.

Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. Sequential latent spaces for modeling the intention during diverse image captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4261–4270, 2019.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 344–354, 2015.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

Melina Aparici, Rocío Cuberos, Naymé Salas, and Elisa Rosado. Linguistic indicators of text quality in analytical texts: Developmental changes and sensitivity to pedagogical work. *Journal for the Study of Education and Development*, 44(1):9–46, 2021.

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. Focus attention: Promoting faithfulness and diversity in summarization. *arXiv preprint arXiv:2105.11921*, 2021.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Eslam Mohamed Bakr, Pengzhan Sun, Li Erran Li, and Mohamed Elhoseiny. ImageCaptioner^2: Image captioner for image captioning bias amplification assessment. *arXiv preprint arXiv:2304.04874*, 2023.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 356–364, 2013.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability With Discourse*, pages 178–186, 2013.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 65–72, 2005.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Eva Banik, Claire Gardent, and Eric Kow. The kbgen challenge. In *Proceedings of the 14th European Workshop on Natural Language Generation (ENLG)*, pages 94–97, 2013.

Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

Anja Belz. Probabilistic generation of weather forecast texts. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 164–171, 2007.

Anja Belz and Ehud Reiter. Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 313–320, 2006.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. MUTAN: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2612–2620, 2017.

Jean-Philippe Bernardy, Shalom Lappin, and Jey Han Lau. The influence of context on sentence acceptability judgements. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 456–461, 2018.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, page 4, 2021.

Anne-Claude Berthoud and Lorenza Mondada. Traitement du topic, processus énonciatifs et séquences conversationnelles. *Cahiers de Linguistique Française*, 17:205–228, 1995.

Shruti Bhargava and David Forsyth. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578*, 2019.

Bin Bi, Chen Wu, Ming Yan, Wei Wang, Jiangnan Xia, and Chenliang Li. Incorporating external knowledge into machine reading for generative question answering. *arXiv preprint arXiv:1909.02745*, 2019.

Derek Bickerton. *Language origins and evolutionary plausibility*. Elsevier Science, 1991.

Kim Binsted and Graeme Ritchie. An implemented model of punning riddles. In *Proceedings of the 12th AAAI Conference on Artificial Intelligence (AAAI)*, pages 633–638, 1994.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, 2009.

Johannes Bjerva, Johan Bos, Rob van der Goot, and Malvina Nissim. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*, pages 642–646, 2014.

Eugen Bleuler and Abraham Arden Brill. *Textbook of psychiatry*. Macmillan, 1924.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT)*, pages 12–58, 2014.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the 1st Conference on Machine Translation (WMT)*, pages 131–198, 2016.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. Findings of the 2017 Conference on Machine Translation. In *Proceedings of the 2nd Conference on Machine Translation (WMT)*, pages 169–214, 2017.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*, 2015.

Johan Bos and Katja Markert. Recognising textual entailment with logical inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP-HLT)*, pages 628–635, 2005.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642, 2015.

Ted Briscoe, Ben Medlock, and Øistein Andersen. Automated assessment of ESOL free text examinations. Technical report, University of Cambridge, Computer Laboratory, 2010.

Gillian Brown and George Yule. *Discourse Analysis*. Cambridge University Press, 1983.

Penelope Brown and Stephen C Levinson. *Politeness: Some Universals In Language Usage*, volume 4. Cambridge University Press, 1987.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer, et al. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 1993.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5016–5026, 2018.

Susanne Burger, Victoria MacLaren, and Hua Yu. The ISL meeting corpus: The impact of meeting type on speech style. In *Proceedings of the International Conference on Spoken Language Processing*, 2002.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. RUBi: Reducing unimodal biases for visual question answering. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Sasha Calhoun, Jean Carletta, Daniel Jurafsky, Malvina Nissim, Mari Ostendorf, and Annie Zaenen. NXT Switchboard annotations. *Linguistic Data Consortium*, 2009.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluation the role of BLEU in machine translation research. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006.

Deborah Cameron. *Verbal Hygiene*. Routledge, 2005.

Shuyang Cao and Lu Wang. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6633–6649, 2021.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018.

Francisco Casacuberta and Enrique Vidal. Giza++: Training of statistical translation models. *Retrieved October*, 29:2019, 2007.

Wallace Chafe. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press, 1994.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568, 2021.

Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. Déja image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 504–514, 2015.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 5935–5941, 2021.

Tao Chen and Min-Yen Kan. Creating a live, public short message service corpus: The NUS SMS corpus. *Language Resources and Evaluation*, 47:299–335, 2013.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. KGPT: Knowledge-grounded pre-training for data-to-text generation. *arXiv preprint arXiv:2010.02307*, 2020.

Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. Acceptability judgements via examining the topology of attention maps. *arXiv preprint arXiv:2205.09630*, 2022.

Colin Cherry and Chris Quirk. Discriminative, syntactic language modeling through latent SVMs. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 65–74, 2008.

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 263–270, 2005.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. *arXiv preprint arXiv:1808.07036*, 2018.

Noam Chomsky. *Syntactic Structures*. De Gruyter Mouton, 1957.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Terry Copeck, Diana Inkpen, Anna Kazantseva, Alistair Kennedy, Darren Kipp, Vivi Nastase, and Stan Szpakowicz. Leveraging DUC. In *Proceedings of the Document Understanding Conferences (DUC)*, 2006.

Terry Copeck, Diana Inkpen, Anna Kazantseva, Alistair Kennedy, Darren Kipp, and Stan Szpakowicz. Catch what you can. *Proceedings of the Document Understanding Conferences (DUC)*, 2007.

Ann Copestake. Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1–9, 2009.

Ann Copestake, Guy Emerson, Michael Wayne Goodman, Matic Horvat, Alexander Kuhnle, and Ewa Muszynska. Resources for building applications with Dependency Minimal Recursion Semantics. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2016.

Courtney Corley and Rada Mihalcea. Measures of text semantic similarity. In *Proceedings of the Empirical Modeling of Semantic Equivalence Workshop at the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005.

Michael A Covington and Joe D McFall. Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100, 2010.

Scott A Crossley. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443, 2020.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer, 2005.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional GAN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2970–2979, 2017.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Edgar Dale and Jeanne S Chall. The concept of readability. *Elementary English*, 26(1): 19–26, 1949.

Robert Dale. Generating recipes: An overview of epicure. *Current Research in Natural Language Generation*, pages 229–255, 1990.

Robert Dale and Chris Mellish. Towards evaluation in natural language generation. In *Proceedings of 1st International Conference on Language Resources and Evaluation*, 1998.

Robert Dale, Barbara Di Eugenio, and Donia Scott. Introduction to the special issue on natural language generation. *Computational Linguistics*, 24(3):345–353, 1998.

Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1574–1583, 2021.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, 2011.

Hoa Trang Dang. DUC 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pages 48–55, 2006.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys*, 51(1):1–40, 2018.

Robert-Alain De Beaugrande and Wolfgang U Dressler. *Introduction to Text Linguistics*, volume 1. longman London, 1981.

Berkan Demirel and Ramazan Gokberk Cinbis. Caption generation on scenes with seen and unseen object categories. *Image and Vision Computing*, 124, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 100–105, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4884–4895, 2019.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Nan Ding, Sebastian Goodman, Fei Sha, and Radu Soricut. Understanding image and text simultaneously: A dual vision-language machine comprehension task. *arXiv e-prints 1612.07833*, 2016.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research*, pages 138–145, 2002.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Mike Dowman, Virginia Savova, Thomas L Griffiths, Konrad P Kording, Joshua B Tenenbaum, and Matthew Purver. A probabilistic model of meetings that combines words and discourse features. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1238–1248, 2008.

John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. Santa Barbara Corpus of Spoken American English. *Linguistic Data Consortium*, 2000.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 489–500, 2018.

Layla El Asri, Hannes Schulz, Shikhar Kr Sarma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: A corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, 2017.

Scott Elliot. Intelli metric: From here to validity. *Automated Essay Scoring: A Cross-disciplinary Perspective*, 2003.

Desmond Elliott and Frank Keller. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 452–457, 2014.

Jeffrey L Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Micha Elsner and Eugene Charniak. Coreference-inspired coherence modeling. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 41–44, 2008.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, et al. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research (JAIR)*, 73:1131–1207, 2022.

Alexander Richard Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-News: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1074–1084, 2019.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4186–4196, 2019.

Youmna Farag, Josef Valvoda, Helen Yannakoudakis, and Ted Briscoe. Analyzing neural discourse coherence models. In *Proceedings of the 1st Workshop on Computational Approaches to Discourse*, pages 102–112, 2020.

Christian Federmann, Oussama Elachqar, and Chris Quirk. Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT)*, pages 17–26, 2019.

Yang Feng, Wanying Xie, Shuhao Gu, Chenze Shao, Wen Zhang, Zhengxin Yang, and Dong Yu. Modeling fluency and faithfulness for diverse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 59–66, 2020.

Dan Flickinger. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28, 2000.

Dan Flickinger. Accuracy vs. robustness in grammar engineering. *Language from a Cognitive Perspective: Grammar, Usage, and Processing*, 201:31–50, 2011.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André FT Martins. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 4963–4974, 2022.

Eric N Forsyth and Craig H Martell. Lexical and discourse analysis of online chat dialog. In *Proceedings of the International Conference on Semantic Computing*, pages 19–26, 2007.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the 2021 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 478–487, 2021.

Yarin Gal and Phil Blunsom. A systematic bayesian treatment of the IBM alignment models. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 969–977, 2013.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 562–569, 2003.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. Grounded response generation task at DSTC7. In *Proceedings of AAAI Dialog System Technology Challenges Workshop*, 2019.

Julia R Galliers and Karen Spärck Jones. Evaluating natural language processing systems. Technical report, University of Cambridge, Computer Laboratory, 1993.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. Generating multiple diverse responses for short-text conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 6383–6390, 2019a.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and William B Dolan. Jointly optimizing diversity and relevance in neural response generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1229–1238, 2019b.

Yang Gao, Wei Zhao, and Steffen Eger. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1347–1354, 2020.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*, pages 124–133, 2017.

John S Garofolo, Christophe Laprun, Martial Michel, Vincent M Stanford, and Elham Tabassi. The NIST meeting room pilot corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2004.

Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research (JAIR)*, 61:65–170, 2018.

Sebastian Gehrmann, Falcon Z Dai, Henry Elder, and Alexander M Rush. End-to-end content and plan selection for data-to-text generation. *arXiv preprint arXiv:1810.04700*, 2018.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*, 2019.

António Góis, Kyunghyun Cho, and André FT Martins. Learning non-monotonic automatic post-editing of translations from human orderings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 205–214, 2020.

Eli Goldberg, Norbert Driedger, and Richard I Kittredge. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53, 1994.

Daena J Goldsmith. Brown and Levinson's politeness theory. *Explaining Communication: Contemporary Theories and Exemplars*, pages 219–236, 2007.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*, pages 166–175, 2019.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards knowledge-grounded open-domain conversations. In *Proceedings of INTERSPEECH*, pages 1891–1895, 2019.

Dionysis Goutsos. *Modeling Discourse Topic: sequential relations and strategies in expository text*, volume 59. Greenwood Publishing Group, 1997.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics (TACL)*, 10: 522–538, 2022.

Tanya Goyal and Greg Durrett. Evaluating factuality in generation with dependency-level entailment. In *Findings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3592–3603, 2020.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6645–6649, 2013.

CC Green, DR Reddy, B Ritea, et al. Speech understanding systems: Report of a steering committee. *Artificial Intelligence*, 9(3):307–316, 1977.

Herbert Paul Grice. Logic and conversation. In *Syntax and semantics: Speech act*, pages 41–58. Elsevier, 1975.

Ralph Grishman and Beth M Sundheim. Design of the MUC-6 evaluation. In *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Virginia*, pages 413–422, 1996.

Anne Grobet. *L'identification des topiques dans les dialogues*. De Boeck Supérieur, 2002.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307, 1993.

Beliz Gunel, Chenguang Zhu, Michael Zeng, and Xuedong Huang. Mind the facts: Knowledge-boosted coherent abstractive text summarization. *arXiv preprint arXiv:2006.15435*, 2020.

Ankush Gupta, Yashaswi Verma, and C Jawahar. Choosing linguistics over vision to describe images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 26, pages 606–612, 2012.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 417–434, 2020.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics (TACL)*, 6:437–450, 2018.

MAK Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, 1976.

Mika Hämäläinen and Khalid Alnajjar. Human evaluation of creative NLG systems: An interdisciplinary survey on recent papers. *arXiv preprint arXiv:2108.00308*, 2021.

Aaron LF Han, Derek F Wong, and Lidia S Chao. LEPOR: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 441–450, 2012.

Seungju Han, Beomsu Kim, and Buru Chang. Measuring and improving semantic diversity of dialogue generation. *arXiv preprint arXiv:2210.05725*, 2022.

David Hardcastle and Donia Scott. Can we evaluate the quality of generated text? In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2008.

Hamza Harkous, Isabel Groves, and Amir Saffari. Have your text and use it too! End-to-end neural data-to-text generation with semantic fidelity. *arXiv preprint arXiv:2004.06577*, 2020.

Helen F Hastie and Anja Belz. A comparative evaluation methodology for NLG in interactive systems. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 4004–4011, 2014.

Basil Hatim and Ian Mason. *The translator as communicator*. Routledge, 2005.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 174–180, 2014.

Johann Gottfried Herder. *Über die neuere deutsche Literatur: Fragmente*, volume 1. Aufbau-Verlag, 1767.

Eli Hinkel and Sandra Fotos. The place of grammar instruction in the second/foreign language curriculum. In *New perspectives on grammar teaching in second language classrooms*, pages 27–44. Routledge, 2001.

Leland E Hinsie and Robert J Campbell. *Psychiatric Dictionary*. Oxford University Press, 1970.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Micah Hodosh and Julia Hockenmaier. Focused evaluation for image description with binary forced-choice tasks. In *Proceedings of the 5th Workshop on Vision and Language*, pages 19–28, 2016.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research (JAIR)*, 47:853–899, 2013.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7856–7870, 2021.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2nd DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, 2022.

Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle, and Montse Maritxalar. Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 1753–1762, 2020.

Chiori Hori, Anoop Cherian, Tim K Marks, and Florian Metze. Audio visual scene-aware dialog track in DSTC8. *DSTC Track Proposal*, 2018.

Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356, 2019.

Mokter Hossain and Ilkka Kauranen. Crowdsourcing: A comprehensive literature review. *Strategic Outsourcing: An International Journal*, 8(1):2–22, 2015.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation (INLG)*, pages 169–182, 2020.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17980–17989, 2022.

Luyang Huang, Lingfei Wu, and Lu Wang. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5094–5107, 2020.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*, 2021.

Kellogg W Hunt. Grammatical structures written at three grade levels. *NCTE Research Report No. 3*, 1965.

John Hutchins. Warren Weaver and the launching of MT. *Early Years In Machine Translation*, page 17, 2000.

Angelina Ivanova, Stephan Oepen, Lilja Øvrelid, and Dan Flickinger. Who did what to whom? A contrastive study of syntacto-semantic dependencies. In *Proceedings of the 6th Linguistic Annotation Workshop*, pages 2–11, 2012.

Angelina Ivanova, Stephan Oepen, and Lilja Øvrelid. Survey on parsing three dependency representations for English. In *Proceedings of the Student Research Workshop at the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 31–37, 2013.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. The ICSI meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, 2003.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2022.

Shaojie Jiang and Maarten de Rijke. Why are sequence-to-sequence models so dull. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 81, 2018.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, 2017.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1601–1611, 2017.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1):81–93, 1938.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should VQA expect them to? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2776–2785, 2021.

Ankit Khare and Manfred Huber. Show, infer and tell: Contextual inference for creative captioning. In *Proceedings of the British Machine Vision Conference (BMVC)*, page 20, 2019.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 199–209, 2017.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Abstractive summarization of Reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2519–2531, 2019.

Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5583–5594, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Paul Kingsbury, Stephanie Strassel, Cynthia McLemore, and Robert McIntyre. CALLHOME American English Transcripts. *Linguistic Data Consortium*, 1997.

W Bradley Knox and Peter Stone. Tamer: Training an agent manually via evaluative reinforcement. In *Proceedings of the IEEE International Conference on Development and Learning*, pages 292–297, 2008.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics (TACL)*, 6: 317–328, 2018.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, et al. Findings of the 2022 Conference on Machine Translation. In *Proceedings of the 7th Conference on Machine Translation (WMT)*, pages 1–45, 2022.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the 1st Workshop on Neural Machine Translation (WMT)*, pages 28–39, 2017.

Philipp Koehn, Franz J Och, and Daniel Marcu. Statistical phrase-based translation. Technical report, University of Southern California, 2003.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180, 2007.

Philipp Koen. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 115–124, 2004.

Ravikumar Kondadadi, Blake Howald, and Frank Schilder. A statistical NLG framework for aggregated planning and realization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1406–1415, 2013.

Ioannis Konstas and Mirella Lapata. Concept-to-text generation via discriminative reranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 369–378, 2012.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 469–473, 2018.

Mahnaz Koupaee and William Yang Wang. WikiHow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*, 2018.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 25, 2012.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*, 2021.

Alexander Kuhnle. Evaluating visually grounded language capabilities using microworlds. Technical report, University of Cambridge, Computer Laboratory, 2020.

Alexander Kuhnle and Ann Copestake. ShapeWorld: A new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*, 2017.

Alexander Kuhnle, Huiyuan Xie, and Ann Copestake. How clever is the FiLM model, and how clever can it be. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 957–966, 2015.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics (TACL)*, 7:453–466, 2019.

Kristopher Kyle and Scott Crossley. The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34:12–24, 2016.

Kristopher Kyle and Scott A Crossley. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786, 2015.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Thomas K Landauer, D Laham, and P W Foltz. Automated scoring and annotation of essays with the intelligent essay assessor. *Automated Essay Scoring: A Cross-disciplinary Perspective*, pages 87–112, 2003.

Jey Han Lau, Alexander Clark, and Shalom Lappin. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1618–1628, 2015.

Jey Han Lau, Alexander Clark, and Shalom Lappin. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5): 1202–1241, 2017.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research (JAIR)*, 67:653–672, 2020.

Batia Laufer and Paul Nation. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3):307–322, 1995.

Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1203–1213, 2016.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. Hallucinations in neural machine translation, 2019.

Geoffrey N Leech. *Principles of pragmatics*. Longman, 1983.

Geoffrey N Leech. *The pragmatics of politeness*. Oxford University Press, 2014.

Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, pages 707–710, 1966.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Computing Research Repository (CoRR)*, 2019.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1430–1441, 2018.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 110–119, 2016.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 986–995, 2017.

Zihao Li. The dark side of ChatGPT: Legal and ethical challenges from stochastic parrots and hallucination. *arXiv preprint arXiv:2304.14347*, 2023.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

Annika Lindh, Robert J Ross, Abhijit Mahalunkar, Giancarlo Salton, and John D Kelleher. Generating diverse and meaningful captions. In *International Conference on Artificial Neural Networks*, pages 176–187. Springer, 2018.

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 923–929, 2016.

Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive GAN for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 33, pages 1028–1035, 2019.

Aishan Liu, Huiyuan Xie, Xianglong Liu, Zixin Yin, and Shunchang Liu. Revisiting audio visual scene-aware dialog. *Neurocomputing*, 496:227–237, 2022.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of

unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2122–2132, 2016.

Fenglin Liu, Xian Wu, Shen Ge, Xuancheng Ren, Wei Fan, Xu Sun, and Yuexian Zou. DimBERT: learning vision-language grounded representations with disentangled multimodal-attention. *ACM Transactions on Knowledge Discovery from Data*, 16(1):1–19, 2021a.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1417–1427, 2020.

Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. Towards faithfulness in open domain table-to-text generation from an entity-centric view. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 13415–13423, 2021b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021c.

Chi-kiu Lo. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the 2nd Conference on Machine Translation (WMT)*, pages 589–597, 2017.

Chi-kiu Lo. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the 4th Conference on Machine Translation (WMT)*, pages 507–513, 2019.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4969–4983, 2020.

John L Locke and Barry Bogin. Language and life history: A new perspective on the development and evolution of human language. *Behavioral and Brain Sciences*, 29(3): 259–280, 2006.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*, 2015.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. Blend: A novel combined MT metric based on direct assessment. In *Proceedings of the 2nd Conference on Machine Translation (WMT)*, pages 598–603, 2017.

François Mairesse, Milica Gasic, Filip Jurcicek, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1552–1561, 2010.

Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 27, 2014.

Elena Manishina. *Data-driven natural language generation using statistical machine translation and discriminative learning*. PhD thesis, Université d'Avignon, 2016.

Diego Marcheggiani and Laura Perez-Beltrachini. Deep graph convolutional encoders for structured data to text generation. *arXiv preprint arXiv:1810.09995*, 2018.

Alexander Mathews, Lexing Xie, and Xuming He. SentiCap: Generating image descriptions with sentiments. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919, 2020.

Michael McCarthy. *Discourse Analysis For Language Teachers*. Cambridge University Press, 1991.

Philip M McCarthy. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. PhD thesis, The University of Memphis, 2005.

Danielle S McNamara and Walter Kintsch. Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, 22(3):247–288, 1996.

Chris Mellish and Robert Dale. Evaluation in the context of natural language generation. *Computer Speech & Language*, 12:349–373, 1998.

Paul Michel and Graham Neubig. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 543–553, 2018.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21263–21272, 2022.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anja Belz, and Aleksandar Savkov. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5739–5754, 2022.

Soichiro Murakami, Sora Tanaka, Masatsugu Hangyo, Hidetaka Kamigaito, Kotaro Funakoshi, Hiroya Takamura, and Manabu Okumura. Generating weather comments from meteorological simulations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1462–1473, 2021.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 344–351, 2007.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Ryo Nakamura, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura. Another diversity-promoting objective function for neural dialogue generation. *arXiv preprint arXiv:1811.08100*, 2018.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 280–290, 2016.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2109–2115, 2016.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1797–1807, 2018.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. ESCAPE: A large-scale synthetic corpus for automatic post-editing. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, 2018.

Jun Ping Ng and Viktoria Abrecht. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1925–1930, 2015.

Ngonga Ngomo. The 9th challenge on question answering over linked data (qald-9). *CEUR Workshop Proceedings: Language*, 7(1):58–64, 2018.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. *CEUR Workshop Proceedings: Choice*, 2640:660, 2016.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2673–2679, 2019.

Jingcheng Niu and Gerald Penn. Grammaticality and language modelling. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 110–119, 2020.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*, 2020.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, 2017.

Jason Obeid and Enamul Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. *arXiv preprint arXiv:2010.09142*, 2020.

OpenAI. ChatGPT. *https://openai.com/blog/chatgpt/*, 2022.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 24, 2011.

Liesl M Osman, Mona I Abdalla, James AG Beattie, Susan J Ross, Ian T Russell, James A Friend, Joseph S Legge, and J Graham Douglas. Reducing hospital admission through computer supported education for asthma patients. *BMJ*, 308(6928):568–571, 1994.

Vishakh Padmakumar and He He. Machine-in-the-loop rewriting for creative image captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 573–586, 2022.

Ellis B Page. The imminence of grading essays by computer. *The Phi Delta Kappan*, 47 (5):238–243, 1966.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The SUN attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108:59–81, 2014.

Adam Pauls and Dan Klein. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 959–968, 2012.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Martha Pennington. Grammar and communication: New directions in theory and practice. *New perspectives on grammar teaching in second language classrooms*, pages 77–98, 2002.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Alvaro Peris and Francisco Casacuberta. A bidirectional recurrent neural language model for machine translation. *Procesamiento del Lenguaje Natural*, pages 109–116, 2015.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018.

Adam Poliak. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the 1st Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, 2020.

Carl Pollard and Ivan A Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, 1994.

Matt Post. Judging grammaticality with tree substitution grammar derivations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 217–222, 2011.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the 3rd Conference on Machine Translation (WMT)*, pages 186–191, 2018.

Ratish Puduppully and Mirella Lapata. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9, 2021.

Matthew Purver. Topic segmentation. *Spoken Language Understanding: Systems For Extracting Semantic Information From Speech*, pages 291–317, 2011.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. URL `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 784–789, 2018.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5370–5381, 2019.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1172–1183, 2021.

John Read. *Assessing Vocabulary*. Cambridge University Press, 2000.

Siva Reddy, Danqi Chen, and Christopher D Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics (TACL)*, 7:249–266, 2019.

Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44 (3):393–401, 2018.

Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, 2000.

Ehud Reiter, Chris Mellish, and John Levine. Automatic generation of technical documentation. *Applied Artificial Intelligence an International Journal*, 9(3):259–287, 1995.

Ehud Reiter, Roma Robertson, and Liesl M Osman. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1-2):41–58, 2003.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2):137–169, 2005.

Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. Investigating pretrained language models for graph-to-text generation. *arXiv preprint arXiv:2007.08426*, 2020.

Brian Richards. Type/token ratios: What do they really tell us? *Journal of Child Language*, 14(2):201–209, 1987.

Alan Ritter, Colin Cherry, and William B Dolan. Unsupervised modeling of Twitter conversations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 172–180, 2010.

Jacques Robin and Kathleen McKeown. Empirically designing and evaluating a new revision-based model for summary generation. *Artificial Intelligence*, 85(1-2):135–179, 1996.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4035–4045, 2018.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics (TACL)*, 8:264–280, 2020.

Vasile Rus and Mihai Lintean. An optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, pages 675–676, 2012.

Amrita Saha, Rahul Aralikatte, Mitesh M Khapra, and Karthik Sankaranarayanan. DuoRC: Towards complex language understanding with paraphrased reading comprehension. *arXiv preprint arXiv:1804.07927*, 2018.

Evan Sandhaus. The New York Times Annotated Corpus. *Linguistic Data Consortium*, 6 (12):e26752, 2008.

B Santorini and A Kroch. The syntax of natural language: An online introduction, 2007.

Nicolas Schrading, Cecilia Ovesdotter Alm, Raymond Ptucha, and Christopher Homan. An analysis of domestic abuse discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2577–2583, 2015.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. QuestEval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*, 2021.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96, 2016.

Samira Shaikh, Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Sarah M Taylor, and Nick Webb. MPC: A multi-party chat corpus for modeling social phenomena in discourse. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2010.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. Long and diverse text generation with planning-based hierarchical variational model. *arXiv preprint arXiv:1908.06605*, 2019.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. FOIL it! Find One mismatch between Image and Language caption. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 255–265, 2017.

Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pages 1598–1608, 2021.

Zhan Shi, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. Toward diverse text generation with inverse reinforcement learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4361–4367, 2018.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the 3rd Conference on Machine Translation (WMT)*, pages 751–758, 2018.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 510–526, 2016.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2021–2030, 2020.

David So, Quoc Le, and Chen Liang. The evolved transformer. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5877–5886, 2019.

Karen Spärck Jones. Towards better NLP system evaluation. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey*, 1994.

C Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Prithvishankar Srinivasan and Santosh Mashetty. SportsBERT. *https://huggingface.co/microsoft/SportsBERT*, 2020.

Somayajulu Sripada, Neil Burnett, Ross Turner, John Mastin, and Dave Evans. A case study: NLG meeting weather industry demand for quality and quantity of textual weather forecasts. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 1–5, 2014.

Katherine Stasaski and Marti A Hearst. Semantic diversity in dialogue with natural language inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 85–98, 2022.

Stephanie Strassel, Jeffrey Conn, Suzanne Wagner Evans, Christopher Cieri, William Labov, and Kazuaki Maeda. SLX corpus of classic sociolinguistic interviews. *Linguistic Data Consortium*, 2003.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. Plan-then-generate: Controlled data-to-text generation via planning. *arXiv preprint arXiv:2108.13740*, 2021.

Yang Sun, Shaonan Tian, and Ming Zhou. Lost in translation: What linguistic measurements best measure text quality of online listings. *Procedia Computer Science*, 199: 1474–1477, 2022.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 3104–3112, 2014.

Yui Suzuki, Tomoyuki Kajiwara, and Mamoru Komachi. Building a non-trivial paraphrase corpus using multiple machine translation systems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 36–42, 2017.

Merrill Swain and Michael Canale. The role of grammar in a communicative approach to second language teaching and testing, 1982.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. Target-guided open-domain conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5624–5634, 2019.

Brian Thompson and Matt Post. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the 5th Conference on Machine Translation (WMT)*, pages 561–570, 2020.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 809–819, 2018.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2944–2953, 2019.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT - Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, 2020.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, 2017.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG)*, pages 355–368, 2019.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67, 2021.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. Measuring the diversity of automatic image descriptions. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1730–1741, 2018.

Emiel van Miltenburg, Chris van der Lee, Thiago Castro Ferreira, and Emiel Krahmer. Evaluation rules! On the use of grammars and rule-based systems for NLG evaluation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 17–27, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.

Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5753–5761, 2017.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 1996.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, 2017.

Georg von der Gabelentz. *Ideen zu einer vergleichenden Syntax: Wort und Satzstellung.* Zeitschrift für Völkerpsychologie und Sprachwissenschaft, 1869.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. Judging grammaticality: Experiments in sentence classification. *Calico Journal*, 26(3):474–490, 2009.

Hongmin Wang. Revisiting challenges in data-to-text generation with fact grounding. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG)*, pages 311–322, 2019.

Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4195–4203, 2019.

Ronald Wardhaugh. *How Conversation Works.* Basil Blackwell, 1985.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics (TACL)*, 7: 625–641, 2019.

Nick Webb, David Benyon, and Preben Hansen. Evaluating human-machine conversation for appropriateness. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, 2010.

Johnny Tian-Zheng Wei, Khiem Pham, Brian Dillon, and Brendan O'Connor. Evaluating syntactic properties of seq2seq output with a broad coverage HPSG: A case study on machine translation. *arXiv preprint arXiv:1809.02035*, 2018.

Joseph Weizenbaum. ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. Towards enhancing faithfulness for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2675–2684, 2020.

Egon Werlich. *A text grammar of English.* Quelle & Meyer, 1976.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2253–2263, 2017.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. TransferTransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45, 2020.

Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. *Second language development in writing: Measures of fluency, accuracy, and complexity.* University of Hawaii Press, 1998.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. A controllable model of grounded response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 14085–14093, 2021.

Stratos Xenouleas, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. SUM-QE: A BERT-based summary quality estimation model. *arXiv preprint arXiv:1909.00578*, 2019.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, 2016.

Huiyuan Xie and Ignacio Iacobacci. Audio visual scene-aware dialog system using dynamic memory networks. In *Proceedings of the DSTC8 Workshop at the AAAI conference on artificial intelligence (AAAI)*, 2020.

Huiyuan Xie, Tom Sherborne, Alexander Kuhnle, and Ann Copestake. Going beneath the surface: Evaluating image captioning for grammaticality, truthfulness and diversity. In *Proceedings of the MetaEval Workshop at the AAAI conference on artificial intelligence (AAAI)*, 2020.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. TIAGE: A benchmark for topic-shift aware dialog modeling. In *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1684–1690, 2021.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. TWEETQA: A social media focused question answering dataset. *arXiv preprint arXiv:1907.06292*, 2019.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3940–3949, 2018.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics (TACL)*, 4:401–415, 2016.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1358–1368, 2018.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. TENER: Adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*, 2019.

Sixing Yan. Disentangled variational topic inference for topic-accurate financial report generation. In *Proceedings of the 4th Workshop on Financial Technology and Natural Language Processing (FinNLP)*. Association for Computational Linguistics, 2022.

Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2013–2018, 2015.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-VQA and text-caption. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8751–8761, 2021.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 180–189, 2011.

Jin-ge Yao, Jianmin Zhang, Xiaojun Wan, and Jianguo Xiao. Content selection for real-time sports news construction from commentary texts. In *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*, 2017a.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6580–6588, 2017b.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4894–4902, 2017c.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659, 2016.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2:67–78, 2014.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1628–1639, 2020a.

Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N Bennett, Nick Craswell, and Saurabh Tiwary. Generic intent representation in web search. In *Proceedings of the ACM Special Interest Group in Information Retrieval (SIGIR)*, pages 65–74, 2019a.

Jianmin Zhang, Jin-ge Yao, and Xiaojun Wan. Towards constructing sports news from live text commentary. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1361–1371, 2016a.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5014–5022, 2016b.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213, 2018a.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*, 2019b.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 31, 2018b.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DialoGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 270–278, 2020b.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D Manning, and Curtis Langlotz. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5108–5120, 2020c.

Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 14830–14840, 2021.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 654–664, 2017.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*, 2019.

Giulio Zhou and Gerasimos Lampouras. Informed sampling for diversity in concept-to-text NLG. *arXiv preprint arXiv:2004.14364*, 2020.

Jianing Zhou and Suma Bhat. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5075–5086, 2021.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 13041–13049, 2020.

Wanzheng Zhu and Suma Bhat. GRUEN for evaluating linguistic quality of generated text. In *Findings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 94–108, 2020.

Appendix A

Templates for generating football highlights in FHiG

A.1 Simple templates

Event type: goal

[time] - goal for [team] by [player].
[player] scored for [team] at [time].
[time] - [team] celebrated as [player] scored.
[team]'s [player] scored at [time].
At [time], [player] scored for [team].
[team] scores! Scored by [player] at [time].

Event type: half time

Half time scores: [home] [home_ht] - [away_ht] [away].
The first half ended [home] [home_ht] - [away_ht] [away].
The first half of the game ended [home] [home_ht] - [away_ht] [away].
At 45 minutes the scores were [home] [home_ht] - [away_ht] [away].
[home] vs [away] - half time score [home_ht] [away_ht].
At half time the scores were [home] [home_ht] - [away_ht] [away].

Event type: full time - one team wins

Full time scores: [home] [home_ft] - [away_ft] [away].
Full time scores: [home] [home_ft] - [away_ft] [away]. [win] wins.
The game ended [home] [home_ft] - [away_ft] [away].
The game ended [home] [home_ft] - [away_ft] [away]. [win] wins.
Final scores of the match: [home] [home_ft] - [away_ft] [away].
[win] wins. Final scores: [home] [home_ft] [away] [away_ft].

Event type: full time - draw

The final score ended drawn at [home] [home_ft] - [away_ft] [away].
The game ended with a [home_ft] - [away_ft] draw for [home] and [away].
And it ends in a draw. [home] [home_ft] [away] [away_ft].
The game finished on a draw. [home] [home_ft] - [away_ft] [away].
At the final whistle it's [home] [home_ft] [away] [away_ft].
A [home_ft] [away_ft] draw for [home] and [away].

Event type: red card

[time] - red card for [player] of [team].
[team]'s [player] received a red card at [time].
At [time] [team]'s [player] received a red card.
Red card for [player] of [team] received at [time].
Red card for [team] at [time]. [player] is sent off.
[player] got red carded at [time] for [team].
[team]'s [player] is given a red card at [time].

A.2 Extended templates

Event type: goal

[time] - goal for [team] by [player].
[player] scored for [team] at [time].
[time] - [team] celebrated as [player] scored.
[team]'s [player] scored at [time].
At [time], [player] scored for [team].
[team] scores! Scored by [player] at [time].
At [time] [player] hit the back of the net for [team].
[player] guided it into the net for [team] at [time].
The net rippled from [player]'s goal for [team] at [time].
Jubilation for [team] as [player] scored at [time].
[player]'s goal for [team] sent the crowd wild at [time].
[player] scored at [time] to lift the spirits of [team].
[player] scored a screamer for [team] at [time].
[team] celebrates as [player] guides the ball into the net at [time].
[player] volleyed it into the top right corner to score for [team] at [time].
[time] - [player] celebrates with the other [team] players as he scores.

Event type: half time

Half time scores: [home] [home_ht] - [away_ht] [away].

The first half ended [home] [home_ht] - [away_ht] [away].

The first half of the game ended [home] [home_ht] - [away_ht] [away].

At 45 minutes the scores were [home] [home_ht] - [away_ht] [away].

[home] vs [away] - half time score [home_ht] [away_ht].

At half time the scores were [home] [home_ht] - [away_ht] [away].

The players headed into the changing room for a break, with the scores at [home] [home_ht] – [away_ht] [away].

The players headed into the dressing room for a break, with the scores at [home] [home_ht] - [away_ht] [away].

The players headed into the changing room for a break, with half time scores at [home] [home_ht] – [away_ht] [away].

The players headed into the dressing room for a break, with half time scores at [home] [home_ht] – [away_ht] [away].

The ref blew the half time whistle, and the players made their way off the pitch. Half time scores: [home] [home_ht] - [away_ht] [away].

At half time there is nothing in it between [home] and [away]. [home_ht] - [away_ht].

The managers have 15 minutes to change the course of this game. The break begins at [home] [home_ht] - [away] [away_ht].

What a half of football. The players deserve a break and head into half time with the scores at [home] [home_ht] - [away_ht] [away].

Event type: final time - one team wins

Full time scores: [home] [home_ft] - [away_ft] [away].

Full time scores: [home] [home_ft] - [away_ft] [away]. [win] wins.

The game ended [home] [home_ft] - [away_ft] [away].

The game ended [home] [home_ft] - [away_ft] [away]. [win] wins.

Final scores of the match: [home] [home_ft] - [away_ft] [away].

[win] wins. Final scores: [home] [home_ft] and [away] [away_ft].

At the final whistle the scores were [home] [home_ft] - [away_ft] [away]. [win] wins.

Full time scores: [home] [home_ft] - [away_ft] [away]. [win] walked away victorious.

The game ended [home] [home_ft] - [away_ft] [away] and [win] walked away victorious.

The final whistle was blown, and the final score ended [home] [home_ft] - [away_ft] [away]. [win] celebrated with the fans.

[win] walk away with all three points. Final scores: [home] [home_ft] - [away_ft] [away].

And it's all over! [win] snatch 3 points to move top of the table. The final results were [home] [home_ft] and [away] [away_ft].

[win] celebrates a hard won victory. [home] [home_ft] - [away_ft] [away].

The final score of a dull match was [home] [home_ft] - [away_ft] [away]. [win] victorious, but only just.

The fans of [win] celebrate as the final whistle is blown. [home] [home_ft] - [away_ft] [away].

Event type: final time - draw

The final score ended drawn at [home] [home_ft] - [away_ft] [away].

The game ended with a [home_ft] - [away_ft] draw for [home] and [away].

And it ends in a draw. [home] [home_ft] [away] [away_ft].

The game finished on a draw. [home] [home_ft] - [away_ft] [away].

At the final whistle it's [home] [home_ft] [away] [away_ft].

A [home_ft] [away_ft] draw for [home] and [away].

Both teams walk away with one point. Full time scores: [home] [home_ft] - [away_ft] [away].

And both teams share the points with the final score [home] [home_ft] [away] [away_ft].

A [home_ft] - [away_ft] draw at 90 minutes. The players of [home] and [away] shake hands and leave the pitch.

A thrilling draw played out by [home] and [away]. Final score: [home_ft] - [away_ft].

A point each for [home] and [away]. We finish with the scores at [home_ft] - [away_ft].

Event type: red card

[time] - red card for [player] of [team].

[team]'s [player] received a red card at [time].

At [time] [team]'s [player] received a red card.

Red card for [player] of [team] received at [time].

Red card for [team] at [time]. [player] is sent off.

[player] got red carded at [time] for [team].

[team]'s [player] is given a red card at [time].

[team]'s [player] headed off the pitch at [time] with a red card.

[team]'s [player] headed off the pitch with a red card at [time].

[time]. [team]'s [player] headed off the pitch with a red card.

The referee pulled a red card out for [team]'s player at [time].

[player] walked off with their head down following a red card at [time]. [team] went down to 10 men.

[player] is adamant that they shouldn't have received a red card for that challenge. They've left [team] with a lot to do.

[team] protests as [player] is sent off for a red card foul at [time].

[player] for [team] receives a red card for a leg-breaking tackle.

At [time] stadium erupted as [team]'s [player] received a red card.

Appendix B

Annotation guidelines for the TIAGE dataset

Here I present the annotation guidelines used for the human annotation process in the TIAGE work.

Task description. chit-chat systems are expected to have the ability to proactively change conversational topics when necessary. For occasions when a chat agent runs out of things to say or the current discussion is starting to get boring, topic shifting is a common tactic to keep the conversation going on. This work aims to model topic-shift phenomenon in open-domain dialogue settings. To achieve this, a new dialogue dataset with topic-shift signals is needed.

Data annotation. For each utterance in a dialog, annotators are asked to decide whether the topic of the conversation changes when transiting from the current utterance to the following response. If there is a topic shift, annotators should label the response with "1", otherwise label it with "0".

In conversations, the response of a speaker to the dialogue context usually falls into one of the following cases (see examples in Table B.1):

(a) Commenting on what the other participant just said (the most common scenario);

(b) Question answering;

(c) Developing the conversation to sub-topics;

(d) Introducing a relevant but different topic;

(e) Completely changing the topic.

Other tips for data labelling. A number of words and phrases are often used as indicators for topic shifts, including but not limited to: but, speaking of, talking about, anyway, by the way, that reminds me, before I forget, I want to mention, let's talk about, we need to discuss, funny you should mention that, not to change the subject but, changing the topic slightly, totally unrelated, on a different/relevant note.

[Speaker1:]	My dad works for the New York Times.
[Speaker2:]	Oh wow! You know, I dabble in photography; maybe you can introduce us sometime.
[Speaker1:]	***Photography is the greatest art out there.*** → not a topic shift

(a) Commenting on the previous context.

[Speaker1:]	Do you teach cooking?
[Speaker2:]	***No, since I'm a native of Mexico, I teach Spanish.*** → not a topic shift

(b) Question answering.

[Speaker1:]	Pets are cute!
[Speaker2:]	***I heard that Huskies are difficult dogs to take care of.*** → not a topic shift

(c) Developing the conversation to sub-topics.

[Speaker1:]	You are an artist? What kind of art, I do American Indian stuff.
[Speaker2:]	Yes, I love to draw. ***I love to eat too, sometimes too much.*** → topic shift

(d) Introducing a relevant but different topic.

[Speaker1:]	What do you do for fun?
[Speaker2:]	I drive trucks so me and my buds go truckin in the mud.
[Speaker1:]	Must be fun! My version of that's running around a library!
[Speaker2:]	***Do you have a favourite animal? Chickens are my favourite. I love them.*** → topic shift

(e) Completely changing the topic.

Table B.1: Different scenarios of dialogue response in conversations.