Event-Related Features in Feedforward Neural Networks Contribute to Identifying Causal Relations in Discourse

Edoardo Maria Ponti
LTL, University of Cambridge
ep490@cam.ac.uk

Anna Korhonen
LTL, University of Cambridge
alk23@cam.ac.uk

Abstract

Causal relations play a key role in information extraction and reasoning. Most of the times, their expression is ambiguous or implicit, i.e. without signals in the text. This makes their identification challenging. We aim to improve their identification by implementing a Feedforward Neural Network with a novel set of features for this task. In particular, these are based on the position of event mentions and the semantics of events and participants. The resulting classifier outperforms strong baselines on two datasets (the Penn Discourse Treebank and the CSTNews corpus) annotated with different schemes and containing examples in two languages, English and Portuguese. This result demonstrates the importance of events for identifying discourse relations.

1 Introduction

The identification of causal and temporal relations is potentially useful to many NLP tasks (Mirza et al., 2014), such as information extraction from narrative texts (e.g., question answering, text summarization, decision support) and reasoning through inference based on a knowledge source (Ovchinnikova et al., 2010).

A number of resources provide examples of causal relations annotated between event mentions (Mirza et al., 2014) or text spans (Bethard et al., 2008). Among the second group, there are corpora compliant with the assumptions of the Rhetorical Structure Theory (RST) in various languages (Carlson et al., 2002; Aleixo and Pardo, 2008), and the Penn Discourse Treebank (Prasad et al., 2007). The latter counts the largest amount of examples and is the only resource distinguishing between explicit and implicit relations.

The discourse signal marking causal relations is often ambiguous (i.e. shared with other kinds of relation), or lacking altogether. Identifying implicit causal relations is challenging for several reasons. They often entail a temporal relation of precedence, but this condition is not mandatory (Bethard et al., 2008; Mirza et al., 2014). Moreover, implicit causal relations are partly subjective and have low inter-annotator agreement (Grivaz, 2012; Dunietz et al., 2015). Finally, they have to be detected through linguistic context and world knowledge: unfortunately, this information cannot be approximated by explicit relations deprived of their signal (Sporleder and Lascarides, 2008). Notwithstanding the partial redundancy between signal and context, implicit examples and explicit examples belonging to the same class appear to be too dissimilar linguistically.

Although various techniques have been proposed for the task, ranging from distributional metrics (Riaz and Girju, 2013, *inter alia*) to traditional machine learning algorithms (Lin et al., 2014, *inter alia*), few have been based on deep learning. Those that have used deep learning have mostly relied on lexical features (Zhang et al., 2015; Zhang and Wang, 2015). The aim of our work is to enrich Artificial Neural Networks with features that capture insights from linguistic theory (§ 2) as well as related works (§ 3). In particular, they capture information about the content and position of the events involved in the relation. After presenting the datasets (§ 4), the method (§ 6) and the experimental results (§ 7) we conclude (§ 8) by highlighting that the observed improvements stem from the link between event semantics and discourse relations. Although our work focuses on implicit causal relations, the proposed features are shown to be beneficial also for explicit instances.

2 Events in Linguistic Theory

Events are complex entities bridging between semantic meaning and the syntactic form (Croft, 2002). The token expressing an event in a text is called a mention and usually consists in a verbal predicate. An event denotes a situation and consists of various components, such as participants and aspect. Participants are entities taking part in the situation, each playing a specific semantic role (Fillmore, 1968; Dowty, 1991). Aspect is the structure of the situation over time and is partly inherent to verbs (Vendler, 1967).

Within discourse, events can establish between themselves different kinds of relation, among which a causal relation (Pustejovsky et al., 2003). This relation is asymmetrical, bridging between a cause and an effect. Discourse-level causation is expressed explicitly through verbs (e.g. *to cause* or *to enable*) (Wolff, 2007) or adverbial markers, either inter-clausal (e.g. *because*) or inter-sentential (e.g. *indeed*). These markers are often ambiguous. Moreover, causation is not necessarily explicit: it can be entailed by the speakers and inferred by the listeners only through world knowledge (Grivaz, 2012).

Both explicit and implicit relations are regulated by a long-standing cognitive principle, namely diagrammatical iconicity. According to this principle the tightness of the morphosyntactic packaging of two expressions is proportional to the degree of semantic integration of the concepts they denote (Haiman, 1985). The relevance of this principle for causal relations has been validated empirically by comparing constructions used to describe causation in visual stimuli (Kita et al., 2010): such constructions were affected by the mediation of an animate participant and the absence of spatial contact or temporal contiguity.

This principle is useful to distinguish causality from other relations. Among adverbial clauses, those expressing cause preserve more independence from the main (effect) clause than the others cross-linguistically. Independence is measured by the freedom in their relative order, the autonomous intonation contour, and non-reduced grammatical categories or valence of verbs (Lakoff, 1984; Diessel and Hetterle, 2011; Cristofaro, 2005). The iconicity principle predicts that this morphosyntactic behaviour corresponds to situations not necessarily sharing time, place and participants from a semantic point of view.

3 Previous Work

Many previous works identified causal relations using metrics or traditional machine learning algorithms. Metrics of the 'causal potential' of event pairs were estimated using distributional information (Beamer and Girju, 2009), verb pairs (Riaz and Girju, 2013) or discourse relation markers (Do et al., 2011). Other techniques employed manually defined rules, consisting in high-level patterns (Grivaz, 2012) or a set of axioms (Ovchinnikova et al., 2010).

The machine learning approaches formulated causal relation identification as a binary classification problem. This problem sometimes involved an intermediate step of discourse marker prediction (Zhou et al., 2010). Features based on fine-grained syntactic representations proved particularly helpful (Wang et al., 2010), and were sometimes supplemented with information about word polarity, verb classes, and discourse context (Pitler et al., 2009; Lin et al., 2014).

Few approaches based on deep learning have been proposed for discourse relation classification so far. Zhang et al. (2015) focused on implicit relations. They introduced a Shallow Convolutional Neural Network that learns exclusively from lexical features. It adopts some strategies to amend the sparseness and imbalance of the dataset, such as a shallow architecture, naive convolutional operations, random under-sampling, and normalization. This approach outperforms baselines based on a Support Vector Machine, a Transductive Support Vector Machine, and a Recursive AutoEncoder.

Moreover, related work on nominal relation classification (Zeng et al., 2014; Zhang and Wang, 2015) showed improvements due to using additional features (neighbours and hypernyms of nouns), as well as measuring the relative distance of each token in a sentence from the target nouns. Although these features are possibly relevant for the identification of causal relations, they have not been investigated for this task before.

4 Datasets

We ran our experiment on two datasets representing different annotation schemes and different languages: the Penn Discourse Treebank in English (Prasad et al., 2007) and the CSTNews corpus in Brazilian Portuguese (Aleixo and Pardo, 2008). The Penn Discourse Treebank was chosen because it distinguishes between explicit and implicit rela-

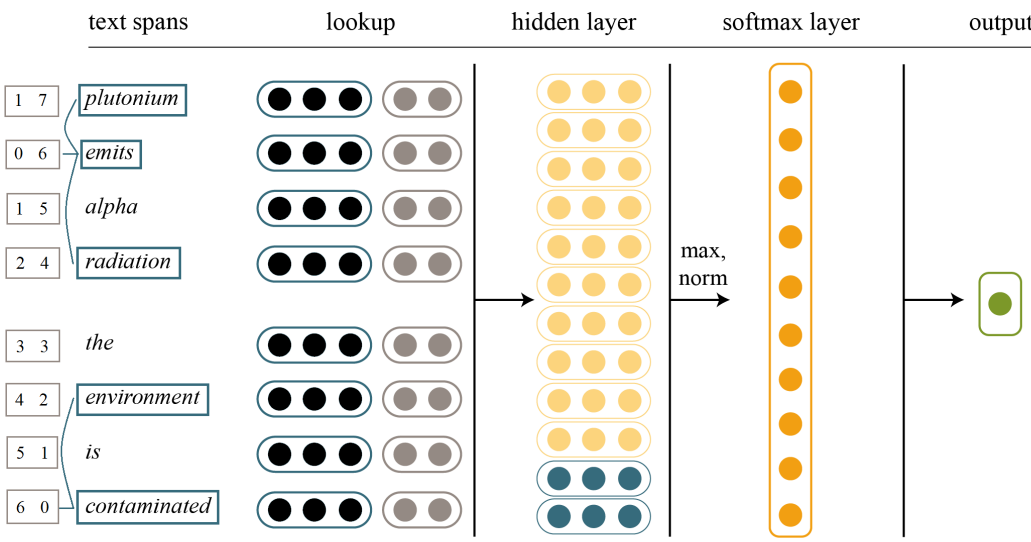

Figure 1: Layers of the Feedforward Neural Network with enriched features.

tions and offers the widest set of examples. Relations are classified into four categories at the coarse-grained level: *Contingency* is considered as the positive class, whereas the others as the negative class.[1] We divided the corpus into a training set (sections 2-20), a validation set (sections 0-1), and a test set (sections 21-22), following Pitler et al. (2009) and Zhang et al. (2015).

On the other hand, he CSTNews corpus contains documents in Brazilian Portuguese annotated according to the Rhetorical Structure Theory. We filtered the texts to keep only relations among leaves in the discourse tree (i.e. containing text spans). The examples labelled as *volitional-cause*, *non-volitional-cause result*, and *purpose* were assigned to the positive class and the others to the negative class. In this case, no distinction was available between implicit and explicit relations. The data partitions in the datasets are detailed in Table 1.

Set	PDTB	CSTNews
Training	3342/9290	190/1101
Validation	295/888	19/143
Test	279/767	19/142

Table 1: Number of examples: positive/negative

5 Features

The most basic kind of features we fed to our algorithm is lexical features, i.e. the vectors stemming from the look-up of the words in every sentence. Vectors are obtained from a model trained with *gensim* (Řehůřek and Sojka, 2010) on Wikipedia. Moreover, we included some additional features: event-related and positional features.

In order to obtain these, the PDTB and CSTNews corpora were parsed using MATE tools (Bohnet, 2010). This parser was trained on the English and Portuguese treebanks available in the Universal Dependency collection (Nivre et al., 2016). In particular, for each of the two related sentences we employed the syntactic trees to discover its root (considered as the event mention) and the nominal modifiers of the root (considered as the participants).[2] We extracted the vector representations of their lemmas, which we call event-related features. Moreover, we assigned to each token two integers representing its absolute linear distance from either event mention. These are called positional features.

The combination of lexical and additional features is called enriched feature set, as opposed to a basic feature set with just lexical features. As an example, consider Figure 1. The lemmas of the two roots are *emit* and *contaminate*. Those of their

[1] *Contingency* overlaps with the fine-grained category *Cause* for implicit relations: *Condition* instead can be hardly conveyed without an explicit hypothetical marker (e.g. *if*).

[2] The syntactic root is often, but not necessarily, a verb. Its nominal modifiers are dependent nouns labelled as subject, direct object, or indirect object.

nominal dependents are *plutonium+radiation* and *environment*, respectively. Moreover, the token *alpha*, for instance, is assigned the integers 1 (distance from *emits*) and 5 (distance from *contaminated*).

The rationale of the additional features is that similar features were employed successfully for nominal relation classification (Zeng et al., 2014). Moreover, they are motivated linguistically. Positional features encode the distance and hence the iconic principle, whereas event-related features account for the semantics of the event and its participants (see § 2).

6 Method

We describe here the architecture of the Feedforward Neural Network with an enriched feature set. The core components of the architecture are a look-up step, a hidden layer and the final logistic regression layer where a softmax estimates the probabilities of the two classes. These are are shown in Figure 1. Positional features are concatenated to the input after the look-up step, and are represented as grey nodes. Event-related features instead are concatenated to the output of the hidden layer, and are represented as blue nodes. The training set was under-sampled randomly: positive examples were pruned in order to obtain the same amount of negative and positive examples.[3] Afterwards, all the sentences of the training set were padded with zeroes to equalize them to a length *n*. Each word was transformed into its corresponding *D*-dimensional vector by looking up a word embedding matrix E. This matrix is a parameter of the model and is initialized with pre-trained vectors. Afterwards, each vector was concatenated along the *D*-dimensional axis with its two neighbouring vectors and its two positional features.

This input representation *x* was then fed to the hidden layer. It underwent a non-linear transformation with a weight and a bias as parameters, and the hyperbolic tangent *tanh* as activation function. The weight is a matrix $W_1 \in \mathbb{R}^{D \times h}$, where *h* is an hyper-parameter defining the size of the hidden layer. The bias, on the other hand, is a vector $b_1 \in \mathbb{R}^h$. Both were initialised by uniformly sampling values from the symmetric interval suggested by Glorot and Bengio (2010). The output of this transformation was concatenated with four word embeddings of the two events and the two (max-pooled) sets of their participants. The resulting matrix underwent a max pooling operation over the *n* axis, which yielded a vector.

Finally, the output of the hidden layer was fed into a Logistic Regression layer. As above, it was multiplied to a weight $W_2 \in \mathbb{R}^{h \times 2}$ and added to a bias $b_2 \in \mathbb{R}^2$. Note that the shape of these parameters along a dimension has length 2 because this is the number of classes to output. Contrary to the hidden layer, both parameters were initialized as zeros. The output of Logistic Regression was squashed by a softmax function σ, which yielded the probability for each class given the example.

The set of parameters of the algorithm is $\theta = \{E, W_1, b_1, W_2, b_2\}$. The loss function is based on binary cross-entropy and is regularised by the squared norm of the parameters scaled by a hyperparameter ℓ. Given an input array of indices to the embedding matrix x_i, the event-related features x_e, the positional features x_p, and a true class y, the objective function is as shown in Equation 1:

$$J = -\sum_{x,y} \sigma(W_2 || \max_n(tanh(W_1 \cdot (x_i \cdot E \oplus x_e) + \\ + b_1) \oplus x_p)|| + b_2) \log P(y) + \ell ||\theta||^2. \quad (1)$$

The optimization of the objective function was performed through mini-batch stochastic gradient descent, running for 150 epochs. Early stopping was enforced to avoid over-fitting. The width of the batches was set to 20, whereas the learning rate λ to 10^{-1}. The vector dimension *D* in the word embedding was 300, the regularization factor ℓ 10^{-4}, and the width of the hidden layer *h* 3000.

7 Results

The performance of the classifier presented in § 6 (named Enriched) was compared with a series of baselines. A naive baseline consists in always guessing the positive class (Positive). A more solid baseline is the state of the art for class-specific identification of implicit relations in the PDTB: the Shallow Convolutional Neural Network (SCNN) by Zhang et al. (2015). The configuration of this algorithm, as mentioned in § 3, includes max pooling, random under-sampling, and normalization. Finally, the last baseline is our

[3] Without random under-sampling, the algorithm worsened its performance, whereas no significant differences were observed with random over-sampling.

Classifier	Macro-F1	Precision	Recall	Accuracy
Positive	42.11	26.67	100	26.67
SCNN	52.04	39.80	75.29	63.00
Basic	53.01	42.04	71.74	**66.44**
Enriched	**54.52**	**42.37**	**76.45**	66.35

Classifier	Macro-F1	Precision	Recall	Accuracy
Positive	21.11	11.80	100	11.80
Basic	48.36	35.51	76.48	82.82
Enriched	**55.62**	**40.66**	**88.24**	**85.00**

Table 2: Different settings for the datasets PDTB (above) and CSTNews (below).

classifier deprived of the additional features (Basic): in other words, it hinges only upon the lexical features.

The results for both the PDTB and CSTNews datasets are presented in Table 2.[4] A McNamar's Chi-Squared test determined the statistical significance of the difference between the classes predicted by Enriched and Basic with $p < 0.05$. The enriched features have a positive impact on precision and recall. This effect is not always observed in accuracy: however, this metric is unreliable due to the high number of negative examples. The improvement on the PDTB is clearly related to implicit examples. From the results on the CSTNews corpus, however, it is safe to gather only that identification of causal relations in general is affected.

8 Conclusion

Drawing upon the semantic theory of events and inspired by work on related tasks, we enriched the feature set previously used for the identification of causal relations. Eventually, this set included lexical, positional, and event-related features. Providing this information to a Feedforward Neural Network, we obtained a series of results. Firstly, our method outperformed earlier approaches and solid baselines on two different datasets and in two different languages, demonstrating the benefit of enriched features. Secondly, our experiment confirmed two theoretical assumptions, namely the iconic principle and the complexity of events. In general, exploiting the theory of event semantics contributed significantly to discourse relation classification, demonstrating that these domains are intertwined to a certain extent.

[4]The results for the CSTNews corpus equals to the average of multiple initializations.

References

Priscila Aleixo and Thiago Alexandre Salgueiro Pardo. 2008. *CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory*. ICMC-USP.

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *Computational Linguistics and Intelligent Text Processing*, pages 430–441. Springer.

Steven Bethard, William J Corvey, Sara Klingenstein, and James H Martin. 2008. Building a corpus of temporal-causal structure. In *Proceedings of LREC'16*, pages 908–915.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd international conference on computational linguistics*, pages 89–97.

Lynn Carlson, Mary Ellen Okurowski, Daniel Marcu, Linguistic Data Consortium, et al. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

Sonia Cristofaro. 2005. *Subordination*. Oxford University Press.

William Croft. 2002. *Typology and universals*. Cambridge University Press.

Holger Diessel and Katja Hetterle. 2011. Causal clauses: A cross-linguistic investigation of their structure, meaning, and use. *Linguistic universals and language variation*, pages 21–52.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, pages 188–196.

Charles Fillmore. 1968. The case for case. In Emmon Bach and Robert Harms, editors, *Universals in linguistic theory*, pages 1–88. Holt, Rinehart & Winston.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, volume 9, pages 249–256.

Cécile Grivaz. 2012. *Automatic extraction of causal knowledge from natural language texts*. Ph.D. thesis, University of Geneva.

John Haiman. 1985. Natural syntax. iconicity and erosion. *Cambridge Studies in Linguistics*, (44):1–285.

Sotaro Kita, NJ Enfield, Jürgen Bohnemeyer, and James Essegbey. 2010. The macro-event property: The segmentation of causal chains. In J. Bohnemeyer and E. Pederson, editors, *Event representation in language and cognition*, pages 43–67. Cambridge University Press.

George Lakoff. 1984. Performative subordinate clauses. In *Annual Meeting of the Berkeley Linguistics Society*, volume 10, pages 472–480.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(02):151–184.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

Ekaterina Ovchinnikova, Laure Vieu, Alessandro Oltramari, Stefano Borgo, and Theodore Alexandrov. 2010. Data-driven and ontological analysis of framenet for natural language reasoning. In *Proceedings of LREC'10*, pages 3157–3162.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 683–691.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The Penn Discourse Treebank 2.0 annotation manual. Technical report.

James Pustejovsky, José M Castano, Robert Ingria, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. Technical report, AAAI.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 21–30.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(03):369–416.

Zeno Vendler. 1967. *Linguistics in philosophy*. Cornell University Press.

WenTing Wang, Jian Su, and Chew Lim Tan. 2010. Kernel based discourse relation recognition with temporal ordering information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 710–719.

Phillip Wolff. 2007. Representing causation. *Journal of experimental psychology: General*, 136(1):82.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *Proccedings of COLING'14*, pages 2335–2344.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1507–1514.