A Distributional Semantic Methodology for Enhanced Search in Historical Records: A Case Study on Smell

Stephen McGregor
LATTICE - CNRS
École normale supérieure / PSL
Université Sorbonne nouvelle Paris 3
1, rue Maurice Arnoux
92120 Montrouge, France
semcgregor@hotmail.com

Barbara McGillivray
The Alan Turing Institute / London, UK
University of Cambridge / Cambridge, UK
bmcgillivray@turing.ac.uk

Abstract

In this paper we present a methodology based on distributional semantic models that can be flexibly adapted to the specific challenges posed by historical texts and that allow users to retrieve semantically relevant text without the need to close-read the documents. We focus on a case study concerned with detecting smell-related sentences in historical medical reports. We demonstrate a process for moving from generic domain label input to a more nuanced evaluation of the semantics of smell in a set of sentences extracted from this corpus, and then develop a machine learning technique for compounding scores on a variety of modelling parameters into more effective classifications.

1 Introduction

With the wealth of historical text collections being digitised as part of various Digital Humanities projects, opportunities arise to develop new approaches to enhancing search of such collections beyond the string or word level. Challenging case studies are often offered by humanities scholars interested in retrieving a set of documents relevant to a particular topic or concept.

Experiments on natural language processing (NLP) techniques, however, have often focused on relatively pristine datasets, providing quantifications of linguistic phenomena in a format conducive to large-scale supervised machine learning methodologies. The consequent trend has been a range of models and systems that achieve impressive results on very particular applications, but which are not obviously generalisable to the type of heterogeneous, unannotated data encountered by, for instance, scholars grappling with large-scale corpora (Faruqui et al., 2016).

In this paper, we explore a methodology grounded in the distributional semantic modelling paradigm specifically designed to deal with data that is potentially dense, messy, and unprocessed. We offer three contributions that will provide a platform for the productive application of NLP techniques to humanities research:

1. A method for extracting useful, nuanced semantic cues from a basic and generic indication of a conceptual domain;

2. A method for using information from distributional semantic models to detect instances where signification of semantic content may be obscure or implicit;

3. An application of a machine learning technique for applying output from our methodology to large-scale data based on annotations of a fraction of the overall corpus.

In order to exemplify the application of our methodology, we will focus on a real-life case study in which a researcher desires to extract sentences from a large historical corpus of medical reports that pertain semantically to smell. An important feature of this particular problem is that smell is often implied rather than explicated in this corpus, as in the following sentence, from the 1913 London County Council Medical Officer of Health report (see section 3 for a description of these reports):

> Foul breath, not depending on carious teeth or obviously septic tonsils, was recorded in a fairly large number of cases, 56 or 31.0 per cent.

For this reason, we take smell terms as a good target for exploring how our methodology can handle not only heterogeneous data, but also ambiguous semantics. While this project serves as a useful illustration of our approach, we maintain that it

should be generalisable to any number of conceptual domains.

The code for this research is available on the GitHub repository https://github.com/BarbaraMcG/Smelly-London/releases/tag/KONVENS-smelly-London (DOI: 10.5281/zenodo.1403213).

2 Background

The distributional semantic paradigm seeks to use observations of word co-occurrences across large scale textual corpora in order to build up mathematically tractable lexical semantic representations that facilitate the projection of words into typically high-dimensional vector spaces in which geometric properties correspond to semantic relationships: so, for instance, words that are close to one another in a distributional semantic space are typically expected to be related in meaning (Clark, 2015). The underlying theoretical assumption of distributional semantics is encapsulated in the *distributional hypothesis* (Harris, 1957), which holds that words that are observed to occur in similar contexts are likely to be related in meaning. Distributional semantic modelling is, crucially, equipped to generally find *paradigmatic* relationships between words that are in some sense and in certain contexts related in meaning, rather than to flesh out the *syntagmatic* relationships inherent in the way that words directly co-occur with one another (Sahlgren, 2008).

While distributional semantic models are often trained on large-scale corpora, the nature of the tasks on which these models have been evaluated has typically, in the tradition of NLP, involved highly structured data. This has entailed, for instance, lists of words rated for degrees of relatedness (Finkelstein et al., 2002) or words and phrases aligned across multiple languages (Klementiev et al., 2012). Applications have included image labelling (Frome et al., 2013; Karpathy and Fei-Fei, 2017) and semantic parsing (Beltagy et al., 2014).

Some recent studies have explored the way that distributional semantic approaches can reliably discover relatively coarse indications of semantic relationships between words, but they are generally not as good at making more specific distinctions regarding, for instance, degrees of synonymy (Hill et al., 2015), suggesting that a distributional semantic model may struggle with capturing the semantic nuance that is often of interest to a scholar approaching a large-scale historical corpus. In these circumstances, it could be very time-consuming or even impossible for the scholar to identify mentions of a given concept via non-explicit semantic cues in the text, for example.

There has been meaningful work experimenting with the way that distributional semantic techniques can generate vector spaces in which geometric properties such as direction can correspond to something that begins to resemble a conceptual space (Mikolov et al., 2013; Derrac and Schockaert, 2015), but the idea of applying these techniques to the type of complex, ambiguous data that is targeted by digital humanists remains relatively unexplored. Complex mathematical techniques have been developed for applying distributional representations to composition (Coecke et al., 2011), but the application of these models to tasks such as word sense disambiguation (Kartsaklis and Sadrzadeh, 2013) and metaphor detection (Gutiérrez et al., 2016) has generally involved training on large sets of manually annotated data. Our objective is to explore a methodology for using distributional semantic techniques to make semantic classifications across messy corpora, learning to generalise from relatively sparse annotations.

In the context of Digital Humanities, Hope and Witmore (2010) have suggested an application for word-counting techniques in the extrapolation of *principal components* from a literary corpus, which could correspond to, for instance, plot elements used to geometrically classify the genres of Shakespeare's plays. Muzny et al. (2017) do not count words; instead they build up distributional representations of parts-of-speech to develop an unsupervised model for classifying dialogism in a sample of text. We seek to contribute to the small but growing literature on applying these types of statistical techniques to problems in making semantic distinctions regarding heterogenous historical data.

3 Data

We focused on British historical medical texts as a case study. The data consists of approximately 5,800 Medical Officer of Health (MOH) reports for London covering the years from 1848 to 1972, which were digitized by the Wellcome Trust in 2012. The MOH reports were published annually by the Medical Officers of Health, who were employed by local authorities in the United Kingdom. These reports contained an overview of the health

of the population at the time, and offer a unique perspective on the everyday lives of Londoners over several generations. Stemming from reaction to infectious disease in mid-19th century, they are important sources for 19th and 20th century history of Public Health in that country.

The MOH reports have been digitally analyzed for the first time in the context of the Smelly London project.[1] Building on the Smelly London project, we decided to focus on the topic of smell as a case study to apply our methodology for enhanced search in the MOH reports.

3.1 Data Cleaning

The data displays a high degree of inconsistency, as is common in the case of historical texts (Piotrowski, 2012), and therefore provides a good test bed for our methodology. In particular, as the MOH reports are the product of contributions from a variety of authors over the course of more than a century, the corpus is characterised by a range of different styles, including text in the format of lists and tables alongside more conventional sentential content. This presents an interesting challenge for both data cleaning and semantic modelling.

In order to process the corpus using distributional semantic methodologies, it is necessary to clean the data and determine sentence boundaries. We apply standard pre-processing measures such as elimination of most punctuation. In a historical corpus such as the one with which we are working, variations and mistake in spelling, not to mention errors in the digitisation process, are not uncommon. That said, we suggest that distributional semantic modelling should be robust to these features of the corpus, smoothing over errors, pushing common variants of word forms close to one another in the output spaces, and perhaps even leveraging very typical errors to gain some semantic traction. We have chosen not to apply lemmatisation here.

Furthermore, the original MOH reports contain a high degree of tabular information, and this has resulted in a good deal of non-sentential strings being picked up in the digitally processed version of the data made available by the Wellcome Collection. In order to address this, we apply a constraint by which any sentence containing a ratio of more than 0.33 non-alphabetical characters to alphabetical characters are thrown out. Figure 1 shows the distribution of these ratios across the corpus: the majority of sentences are concentrated around the 0.2 point, indicating that most strings identified as complete sentences contain mainly alphabetic characters. At the same time, this culling technique results in the elimination of a long tail of undesirable data, including a number of inordinately long sentences that might perturb our word counts.

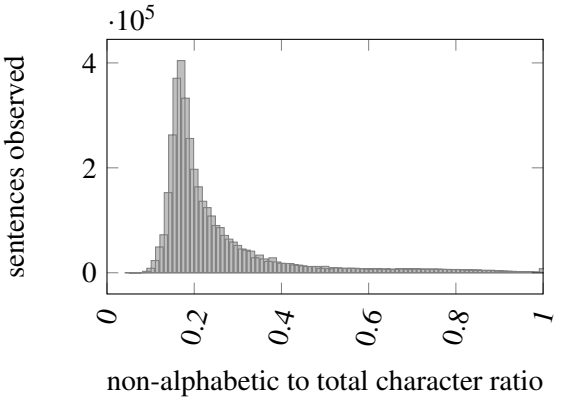

Figure 1: Histogram of ratios of non-alphabetic to total character count for sentences in the Medical Officer of Health corpus.

In addition to the above measures, we use generic encodings to represent years, ranks, and any other string of digits. This step is intended to provide a semantic handle for words that tend to occur in proximity to a variety of different instances of any of these numeric types. This data cleaning process results in a pared down corpus consisting of 3,195,696 sentences containing 81,552,482 word tokens representing 388,320 word types.

4 Methodology

In this section, we will outline the process by which we create a computational framework for enhancing search in historical texts and apply it to the task of identifying sentences that indicate smell. This modelling procedure encompasses a range of unsupervised techniques, working from the premise that a combination of distributional semantic applications and sentence-level heuristics can give us a handle on how we might meaningfully assign interpretations to sentences extracted from the large-scale corpus described above. Our methodology involves three steps:

1. Building a distributional semantic model based on an analysis of our corpus;

[1]The project website url is `https://smelly-london.github.io/Smelly-London/visualisation/leaflet/`

2. Finding candidate smell words in a distributional semantic space;

3. Using our candidate words to make guesses about whether a sentence indicates smell.

Combined, these steps entail a set of parameters, and these parameters will become the basis for the supervised training of an effective multi-variable model that learns to classify sentences in terms of their either implicitly or explicitly connoting smell.

4.1 Distributional Semantic Models

As described in Section 2, the premise of the distributional semantic approach is that words with similar co-occurrence profiles should be semantically and conceptually related. An essential parameter of a distributional semantic model is the window within which words will be considered to co-occur. We use two different techniques for building distributional semantic models, and both involve generating representations from observations across our corpus of words that occur within two words on either side of a vocabulary word, within five words on either side of a vocabulary word, and anywhere in the same sentence as a vocabulary word. We construct our vocabulary from the 60,000 most frequent word types in our corpus, which means that each vocabulary word is observed at least 10 times.

The first of our two modelling techniques involves using a neural network to generate *word embeddings*, which are abstract vector-space representations of vocabulary words iteratively generated from observations made across a series of iterations over our corpus. In particular we apply both of the methods encompassed by the widely discussed `word2vec` technique described by Mikolov et al. (2013).[2] The CBOW (contextual bag-of-words) procedure involves trying to predict a word based on the other words found in proximity to it, while the SKIP-GRAM procedure involves trying to predict the sequence of words within a co-occurrence window around a target word. In both cases, models make use of an arbitrary number of dimensions in a developing vector-space as handles for gradually pulling word-vectors into place.

The second technique we use is the *dynamically contextual* methodology originally described by (McGregor et al., 2015), and subsequently applied to tasks such as metaphor detection (Agres et al., 2016) and semantic coercion classification (McGregor et al., 2017). This approach builds up a statistical representation of each word in our vocabulary over a single traversal of our corpus, using a mutual information weighting to measure the unexpectedness inherent in the observation of two words co-occurring (with co-occurrence defined as above). This results in a very high-dimensional (with, in effect, one dimension for every word type in the corpus, since any word can in principle co-occur with any other word) and very sparse (since most words never will actually co-occur with one another) co-occurrence matrix. This matrix then serves as the basis for *context-specific* projections into lower-dimensional subspaces, which will be described in the following section. These subspaces furthermore afford two different modes of exploration, which will likewise be discussed below.

Baroni et al. (2014) have characterised these two different approaches in terms of *predicting* and *counting* respectively, determining that the predictive embeddings ultimately result in more productive models.[3] Levy and Goldberg (2014) have offered an alternative viewpoint, making the case that both techniques are effectively accomplishing the same thing, with differences in performance corresponding to the tuning of model parameters.

While the statistics that populate the matrix we use for generating conceptual projections are broadly in line with tradition count-based approaches (with measures in place to ensure non-negative values and to avoid selecting obscure co-occurrence features), the dynamic nature of the projected subspaces ensures that the relationship between word-vectors will vary considerably from context to context, a unique feature of this approach. One of our empirical objectives is to compare as well as to combine these different approaches in their application to our novel task of identifying semantic indications of smell, and so each of the modelling techniques will provide the basis for a subset of our overall methodological parameters.

4.2 Extracting Candidate Smell Words

With a few different distributional semantic spaces built, our next task is to use these spaces to generate candidate smell words. Our objective is not simply to discover words that are unambiguously associ-

[2] For this we use the `gensim` library for Python.

[3] In practice, count-based matrices of co-occurrence statistics are often factorised into less interpretable representations (Deerwester et al., 1990), but this does not apply to our contextual application of distributional semantics.

ated with smell – we could simply use a thesaurus to do this – but rather to find words that sometimes, in certain contexts, indicate smell. For this reason, we begin with sets of words that are canonically associated with smell, based on their classification in the WordNet lexical taxonomy (Fellbaum, 1998). We work with sets of multiple seed words; an analysis of a single word could be productive, but might also admit elements of ambiguity associated with that word (*smell* can be used as a verb, for instance). On the other hand, using too many words could result in the selection of overly generic co-occurrence features in our contextual models.

We take as seeds two different WordNet *synsets* (collections of semantic descriptors associated with a particular denotation). From the synset *olfactory_property.n.01*, defined as "any property detected by the olfactory system", we extract the lemmas *smell, aroma, odour,* and *scent* (we ignore words that aren't included in our distributional semantic model vocabulary). This set of words is presumed to correspond to general signifiers of smell. Alternately, starting from the synset *smell.n.01*, defined as "the sensation that results when olfactory receptors in the nose are stimulated by particular chemicals in gaseous form", we take the four synsets that are direct hyponyms within the WordNet taxonomy (*malodour.n.01, acridity.n.01, aroma.n.02,* and *scent.n.02*) and then extract all unique, in-vocabulary lemmas associated with those synsets: *malodor, malodour, stench, stink, reek, fetor, mephitis, acridity, aroma, fragrance, perfume,* and *scent.* This second set of words is taken to represent more specific instances of types of smells (and it is interesting to note that most, but not all, seem to have negative connotations).

It must be noted that, while these words are arguably, because of their relationship to the WordNet knowledge base, canonically about smell, they are not necessarily prime indicators of smell in our corpus. In fact, in the subset of 56,033 words across 1,954 sentences annotated by a human for indications of smell described in Section 5, only one of these canonical smell words (*scent*) is ever observed. This sparsity motivates our semantic modelling procedures as an attempt to move from a representation of smell that is in some sense objective to techniques for extracting specific and contextual instances of smell. The application of the two sets of terms to our distributional semantic models rests upon the distributional hypothesis itself, which suggests that words that tend to occur in similar contexts to our input words, and so are close to them in a distributional semantic space, might also *at least sometimes* be similar in meaning. A key question that we will explore here is whether we can capture something of the context-specific way in which smell is sometimes implied.

In the case of the `word2vec` models, we take each set of input words and find the centroid of the word-vectors corresponding to each of those words, exploring spaces of 20 and 200 dimensions. We then explore the space around this centroid, returning the closest words to the centroid up to a point, with the number of words selected becoming another model parameter: we consider the top 20, 50, 100, and 200 words in our experiments.

With our context-sensitive distributional semantic models, we use the two sets of input words as the basis for projecting subspaces that we hypothesise will correspond to the concept of smell. We apply two different techniques for selecting dimensions from our large, sparse base matrix, in each case using an analysis of the word-vectors associated with the input words to find co-occurrence terms that are expected to be collectively in some sense salient to the input and therefore to the concept that those input words likewise collectively represent:

JOINT Choose the dimensions that have non-zero values for all input words and then have the top mean co-occurrence weights across all words;

INDY Concatenate a subset of dimensions that have the highest value for each input word independently.

We can choose how many top dimensions we return using each of these techniques. In line with the dimensionalities of the `word2vec` spaces, we explore subspaces with 20 and 200 dimensions in the experiments reported below. Unlike with the `word2vec` spaces, each dimension in a contextual projection corresponds to a co-occurrence features, and so to a term that is expected to be in some sense related with the concept indicated by the seed words used as input.

A significant property of these contextualised subspaces is that they have geometric features that spaces of word embeddings generally do not have. Of particular note for the present investigation is the idea of distance from the origin (word embeddings are taken as normalised). Co-occurrence

weights of 0 indicate that the corresponding word pairs are never observed to co-occur with one another, and so word-vectors found at the origin of a subspace are taken to have no conceptual intersection with the input words used to generate the co-occurrence dimensions delineating that subspace.

Conversely, in a contextualised subspace composed of conceptually salient dimensions (ie, dimensions corresponding to co-occurrence terms associated with smell words), we expect word-vectors that have high values across many dimensions and that are correspondingly far from the origin to be likewise closely associated with the concept we are modelling. With this in mind, we use two different techniques to explore our contextualised subspaces: as with the general `word2vec` semantic spaces, we search for words which are close to the centroid of our input words in the subspace (corresponding to the *dist.* parameters in Table 1), and then we also search for words with the largest norm (corresponding to the *norm* parameters). Again in line with the neural network spaces, we find the top 20, 50, 100, and 200 words associated with each of these search techniques.

Here are two examples of the top 10 words output using two different combinations of the parameters described thus far:

Specific input, SKIP-GRAM model, 2x2 window, 20 dimensions: *suffocating, sporules, thread, stance, fibres, core, fibrinous, sacking, cuts, bruising*

Generic input, INDY-norm model, 5x5 window, 200 dimensions: *taste, faint, flavour, odours, disagreeable, colour, sour, unpleasant, pungent, musty*

There are peculiarities associated with each set of parameters to be noted. For instance, the specific input, along with the smaller co-occurrence window, discovers some seemingly off-topic but also perhaps topically coherent words such as *thread*, *fibre*, and *sacking*. The generic input, applied to a contextually dynamic model, on the other hand discovers some words related to smell in a more general way, but also moves into some other sense domains by way of *taste* and *colour*. The hypothesis we seek to test is that, either independently or collectively, the various lists of words associated with iterations of our modelling parameters will prove useful in detecting sentences where smell may be entailed either implicitly or explicitly.

parameter	features	count
inputs	general, specific	2
models	SKIP-GRAM, CBOW, JOINT (dist.), INDY (dist), JOINT (norm), INDY (norm)	6
window	2, 5, full-sentence	3
dimension	20, 200	2
words output	20, 50, 100, 200	4
classification	1-word, 2-words, 0.05-ratio, 0.1-ratio, dependency	5
total		1,440

Table 1: The parameter space of our methodology.

4.3 Identifying Sentences

The process outlined above of building and then exploring distributional semantic spaces generates lists of words which may be associated with connotations of smell. Given a list of candidate smell words, we propose five different techniques for making a positive classification of a sentence:

1. The sentence contains at least one smell word;
2. The sentence contains at least two smell words;
3. The ratio of smell words to total words in a sentence is at least 0.05;
4. The smell-to-total word ratio in a sentence is at least 0.1;
5. There is at least one pair of smell words involved in a dependency relationship, such that one word is a child of the other word in the parse tree of a sentence.

The first four techniques involve straightforward word counting across a sentence. The fifth technique, which entails applying a parser to a sentence being analysed,[4] is intended to introduce an element of compositionality to our methodology.

5 Results and Evaluation

The various techniques for model building, space searching, and sentence classification described throughout the previous section combine to define the parameter space of our methodology, summarised in Table 1. To review, the first four parameters pertain to model building, while the fifth parameter (word output) refers to the number of words

[4]For this purpose we employ Python's `spacy 2.0.11` parser: `https://spacy.io/`

returned from any given distributional semantic space, and the classification parameter refers to the ways in which these lists of candidate smell words are applied to actual sentences.

We evaluated our smell detection system against a gold standard based on human annotation. The medical historian in the team manually annotated 1,954 sentences extracted in random order from four likewise randomly selected MOH reports (Bermondsey 1924, Chelsea 1920, Deptford 1902, and Port of London 1890). The task consisted in assigning a binary label corresponding to whether or not the sentence entailed smell.

5.1 Single Feature Extraction

We compared the gold standard rating for each of the 1,954 annotated sentences against each of the 1,440 combinations of modelling parameters. For each combination we calculated precision and recall as the ratio between the number of correctly tagged sentences and the number of tagged sentences, and the ratio between the number of correctly tagged sentences and the number of sentences in the gold standard, respectively.

The distribution of precision and recall scores across all parameter combinations is very skewed, with most combinations consequently having low F1 scores (see Figure 2 for a visualisation). 52 combinations had a precision score of 1, but very low recall (< 0.020); 21 combinations had a recall score of at least 0.900 (with a maximum of 0.973) but a very low precision (< 0.080). The combination with the highest F1 score (0.235) had precision of 0.194 and recall is 0.297. This combination uses lemmas associated with the WordNet synset generically describing smell as input (general input in Table 1), uses a dynamically contextual technique to select dimensions that are independently salient to each input word while returning the words that are closest to the centroid of input words (INDY-dist), considers the entire sentence as a co-occurring bag-of-words (full-sentence window), projects into a 200-dimensional space, outputs 200 words, and judges a sentence as being about smell if two output words are in it (2-words classification).

This top performing parameter combination is to a certain extent exemplary. In fact, of the top 50 feature combinations in terms of F1 score (out of a total of 656 combinations with non-zero F1 values), 46 involve the INDY technique for selecting conceptually contextualised subspaces. It would

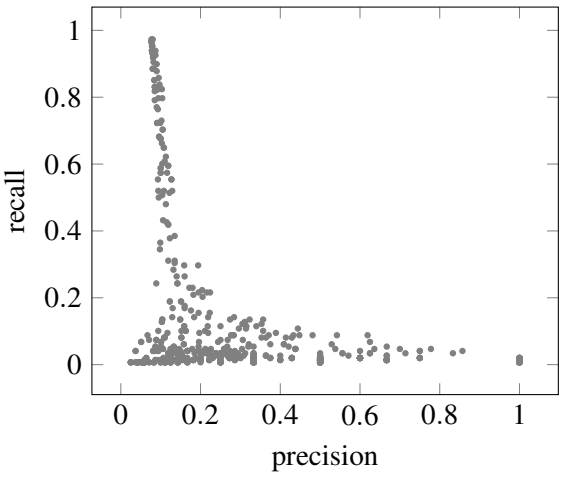

Figure 2: Precision and recall for each of the parameter combinations with at least one positive identification of a sentence that is about smell.

seem, then, that the information contained in dimensions independently associated with each of a set of smell-related input words does a good job of specifying a subspace where the semantics of smell take on a predictable geometry.

Some of the parameter combinations returned a very high number of positive classifications. The most permissive modelling set-up identified 1,870 out of 1,954 sentences as being about smell, and there were 27 parameter combinations that predicted at least 75% of sentences were about smell. These highly permissive versions of our methodology are most notably drawn entirely from dynamically contextual distributional semantic modelling techniques utilising the distance metric. The suggestion here is that this technique tends to discover highly generic words likely to be found in many different sentences—but less likely to signal the semantics of smell in particular. The more specialised set of input words extracted from WordNet also feature strongly (in 19 out of the 27), and this again highlights an aspect of distributional semantic modelling: as more words are used to define either a region or a subspace, the specified geometry tends to pertain to increasingly generic terms.

These results must be considered in the context of the performance of a minority class baseline: if we were to guess that every sentence in our dataset were about smell, we would achieve an F1 score of 0.141, and only 140 out of the 656 feature combinations that return non-zero F1 scores would do better than this. In other words, using these mechanisms for statistical analysis of models consisting of a

single combination of parameters, there will be a tendency to identify those models with high recall and at least some precision—which is to say, those models that tend towards identifying everything as being about smell. This approach of returning almost every input sentence would presumably be of little value to a scholar faced with the task of reviewing a massive collection of text.

5.2 Multiple Feature Evaluation

Clearly, a single set of smell-related words is not sufficient to reliably indicate sentences that are about smell. With this in mind, we next consider the possibility of learning a way to analyse results from a multitude of model parameter combinations that might give us a better chance of classifying the semantics of a sentence. The objective of this experiment is to apply a machine learning approach to our data in hopes of discovering more complex patterns across sets of lists of words which, on their own, come up short of being comprehensively indicative of the semantics we hope to extract.

We begin by seeding a logistic regression with the sentence-by-sentence classifications for the top scoring combination of parameters in terms of precision, recall, or F1 score. Moving down the list of ranked scores for a given metric, we add to our regression the next best performing set of classifications with the constraint that the coefficient of determination associated with adding this next set to the established set is no greater than 0.5. Formally, given an established matrix X of k parameter combination classifications over n sentences, we admit a column of classifications x associated with candidate combination x_{k+1} for inclusion based on the following constraint, where β is the vector of k coefficients learned in a least mean squares linear regression treating x_{k+1} as the dependent variable:

$$y_i = \sum_{j=1}^{k} \beta_j \times X_{i,j} \tag{1}$$

$$1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\mathrm{Var}(y_i)} < 0.5 \tag{2}$$

This is equivalent to setting a *variance inflation factor* of 2, and is intended to inhibit our regression from acquiring features with rampant collinearity (O'Brien, 2007). In practice, because our independent variables are in fact just lists of ones and zeros, for a given sentence we are effectively just taking the sum of coefficients that correspond to positive classifications for that sentence and then taking either the square of that sum or the square of one minus that sum, depending on whether a new parameter combination would guess that sentence has to do with smell or not, then dividing the sum of all these sums across all sentences by the variance of the sums of coefficients for positive classification parameter combinations for each sentence.

seed	prec.	rec.	F1
precision	0.800	0.324	0.460
recall	0.758	0.340	0.469
F1 score	0.770	0.362	0.492
minority class	0.076	1.000	0.141

Table 2: Results for multi-feature logistic regressions on sets of parameters seeded with top single-feature results for each statistical metric, constrained by a variance inflation factor.

This technique results in the generation of a set of 155 non-collinear parameter combinations seeded with the best precision results, a 162 combination set for recall results, and 159 combinations for F1 scores. Each of these sets of parameter combinations can be conceived of as a binary matrix, with one column for each combination of parameters that was included based on the variance inflation factor constraint and one row for each sentence in our datasets. Each of these matrices then becomes the input for a logistic regression model, in which the columns of scores assigned to sentences by parameter combinations are independent variables and the regression attempts to learn to predict the human assigned scores for each sentence.[5]

Results are reported in Table 2, along with a minority class baseline. The top performing F1 Score of 0.492, which, not surprisingly, is achieved by the matrix of parameters seeded with the top single-combination features for F1 Scores, is significantly stronger than the 0.235 achieved by the best single-combination feature itself ($p < .02$, based on a permutation test), not to mention the minority class baseline. It is notable that all three of the seeding strategies result in considerably higher precision than recall, indicating that our regression is in all cases learning to error on the side of caution, so to speak, making relatively few classifications with a relatively high degree of accuracy.

These logistic regressions can be further anal-

[5] We build our logistic regression using the `sklearn` module for Python.

ysed by considering the specific coefficients assigned to each parameter combination feature of each input matrix. What is most notable about the top most positive features for the regression seeded with top F1 scores is the diversity of parameters represented: across the top ten features, there is an even split of general and specific WordNet input; combinations involving both static and dynamic distributional semantic models, and both the norm and distance techniques amongst the dynamic models; multiple instances of each of the three co-occurrence window values; all four values for total number of candidate words; and four of the five sentence classification techniques. Amongst the ten parameter combinations assigned the most negative coefficients, the one thing that stands out is a propensity for 20 dimensional distributional semantic spaces over 200 dimensional spaces (9 out of the 10), suggesting that these lower dimensional spaces are more likely to come up with irrelevant candidate smell words that end up becoming counterindications of the semantics of smell when they are observed in a sentence.

6 Conclusion

We have presented a general methodology to enhance search in historical texts by using distributional semantic models that help find instances of words related to a specific concept. In particular, we developed a method for extracting semantic cues from a basic and generic indication of the conceptual domain of smell (section 4.2) and a method for detecting implicit instances of the smell concept (section 4.3). We substantially enhanced our methodology's output by applying a machine learning technique designed to discover patterns between different sets of words generated by varying our modelling parameters. We applied this methodology to the detection of smell-related sentences in a corpus of historical medical records. In the future we plan to test the methodology on other historical texts and other concepts. For the time being, the case study presented here should offer a clear picture of how this approach can be generalised to any number of topics.

There are a few variations on our methodology that might be worth considering. For instance, while distributional semantic modelling is geared towards discovering paradigmatic relationships, we might consider accepting the salient co-occurrence terms that delineate our contextualised projections as candidate words in themselves. More ambitiously, we might consider applying recent work on semantic embeddings targeting language at the level of phrases or sentences into our modelling procedure (Le and Mikolov, 2014) (though there might be issues with corpus size here, and a transfer learning approach might be inappropriate for a historical corpus). Finally, an analysis of the way that sets of words associated with different modelling parameters overlap and diverge could provide further evidence for future enhancements.

In terms of expanding our methodology, we might imagine a situation where a researcher is interested in making sure all instances where a particular concept is mentioned are identified, or, conversely, where there is a desire for a high degree of certainty that all identified instances are accurate. Questions like these, which should be examined in collaboration with scholars dealing with large-scale corpora, will motivate further investigation of the way that modelling parameters might be used to weight either precision or recall of the output.

We have also refrained here from considering interesting questions regarding phenomena such as diachronic semantic change. This is a complex topic, but, rather than see the way that word meaning might shift across the scope of a historic corpus as a problem, we imagine that the general methodology we have outlined here might, with further development, provide a mechanism for charting semantic evolution.

Authors' Contributions

S.McG. participated in the design of the study, processed the data, implemented the distributional analysis system, and drafted the manuscript. B.McG. conceived the study, participated in the design of the study, implemented the evaluation, and drafted the manuscript.

Acknowledgements

We would like to thank Deborah Leem for introducing us to this dataset, and for providing the historical research questions, as well as the annotation for the evaluation of the system. This work was supported by the Chist-ERA Atlantis project. This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Kat Agres, Stephen McGregor, Karolina Rataj, Matthew Purver, and Geraint A. Wiggins. 2016. Modeling metaphor perception with distributional semantics vector space models. In *Workshop on Computational Creativity, Concept Invention, and General Intelligence*, 08/2016.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, pages 238–247.

Islam Beltagy, Katrin Erk, and Raymond Mooney. 2014. Semantic parsing using distributional semantics and probabilistic logic. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 7–11, June.

Stephen Clark. 2015. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, pages 493–522. Wiley-Blackwell.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2011. Mathematical foundations for a compositional distributed model of meaning. *Linguistic Analysis*, 36(1-4):345–384.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal for the American Society for Information Science*, 41(6):391–407.

Joaquín Derrac and Steven Schockaert. 2015. Inducing semantic relations from conceptual spaces: A data-driven approach to plausible reasoning. *Artificial Intelligence*, 228:66–94.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transanction on Information Systems*, 20(1):116–131.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129.

E. Darío Gutiérrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin K. Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Zellig Harris. 1957. Co-occurrence and transformation in linguistic structure. *Language*, 33(3):283–340.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with genuine similarity estimation. *Computational Linguistics*, 41(4):665–695.

Jonathan Hope and Michael Witmore. 2010. The Hundredth Psalm to the tune of "Green Sleevess": Digital approaches to Shakespeare's language of genre. *Shakespeare Quarterly*, 61(3):357–390.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2013. Prior disambiguation of word tensors for constructing sentence vectors. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1590–1601.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Bombay, India.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, pages II–1188–II–1196. JMLR.org.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.

Stephen McGregor, Kat Agres, Matthew Purver, and Geraint Wiggins. 2015. From distributional semantics to conceptual spaces: A novel computational method for concept creation. *Journal of Artificial General Intelligence*, 6(1):55–89.

Stephen McGregor, Elisabetta Jezek, Matthew Purver, and Geraint Wiggins. 2017. A geometric method for detecting semantic coercion. In *Proceedings of 12th International Workshop on Computational Semantics*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop*.

Grace Muzny, Mark Algee-Hewitt, and Dan Jurafsky. 2017. Dialogism in the novel: A computational model of the dialogic nature of narration and quotations. *Digital Scholarship in the Humanities*, 32(suppl_2):ii31–ii52.

Robert M. O'Brien. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Morgan & Claypool Publishers.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.