Navigating Literary Text with Word Embeddings and Semantic Lexicons

Susan Leavy[1], Karen Wade[2], Gerardine Meaney[2], and Derek Greene[1]

[2]Humanities Institute, University College Dublin, Ireland
[1]Insight Centre for Data Analytics, University College Dublin, Ireland
{*susan.leavy,karen.wade,gerardine.meaney,derek.greene*}*@ucd.ie*

Abstract

Word embeddings represent a powerful tool for mining the vocabularies of literary and historical text. However, there is little research demonstrating appropriate strategies for representing text and setting parameters, when constructing embedding models within a digital humanities context. In this paper we examine the effects of these choices using a case study involving 18th and 19th century texts from the British Library. The study demonstrates the importance of examining implicit assumptions around default strategies, when using embeddings with literary texts and highlights the potential of quantitative analysis to inform critical analysis.

1 Introduction

This research is part of a digital humanities project exploring attitudes towards disease and illness in the 18th and 19th centuries. The associated corpus contains a large, diverse selection of digitised texts. Lexicons generated using word embeddings are part of a suite of big data approaches which are applied in order to navigate this corpus, which consists of over 46,000 texts dedicated to a range of subjects. It is hoped that these techniques will allow the identification of key texts and thematic trends concerned with illness and disease, so that these can be interpreted with reference to current and historical debates surrounding biopolitics, medical culture, and migration.

Word embeddings are increasingly being used to generate semantic lexicons for a variety of tasks (Mikolov et al., 2013). This includes uncovering changes in the sense of terms over time (Hamilton et al., 2016), extracting social networks from literary texts (Wohlgenannt et al., 2016), and text classification (Leavy et al., 2017). However, there is a lack of research demonstrating optimal strategies for setting parameters, when constructing these models on literary and historical texts. There has also been little study on the effect of text preprocessing decisions on the resulting models (Lapesa and Evert, 2014; Camacho-Collados and Pilehvar, 2017). Given that the assumptions behind preprocessing can have particular significance within a digital humanities context, it is important to explore the impact of these decisions, which is often not reported or considered (Sculley and Pasanek, 2008).

This research evaluates the setting of parameters in word embedding along with standard preprocessing approaches including conversion of all letters to lowercase and removal of stop-words. Evaluation of lexicons generated using word embedding is commonly conducted using an intrinsic approach whereby the resulting lexicons are evaluated against existing standard lexical databases such as WordNet (Miller, 1995). However, given the domain specificity and historical nature of this corpus, extrinsic evaluation was conducted based on the effectiveness of each lexicon in identifying texts that relate to medical topics.

2 Methodology

The corpus used in this research is comprised of a diverse collection of digital texts from the British Library. In the analysis described here, we focus on a subset of 35,916 English language texts, dating from 1700 to 1899. Word2vec Continuous Bag-of-Words (CBOW) and Skip-Gram (SG) embedding models were generated from the English corpus using 30 combinations of parameters and text processing strategies (see Table 1).

A set of 10 seed terms was derived from a 19th century medical reference book (Guy, 1856),

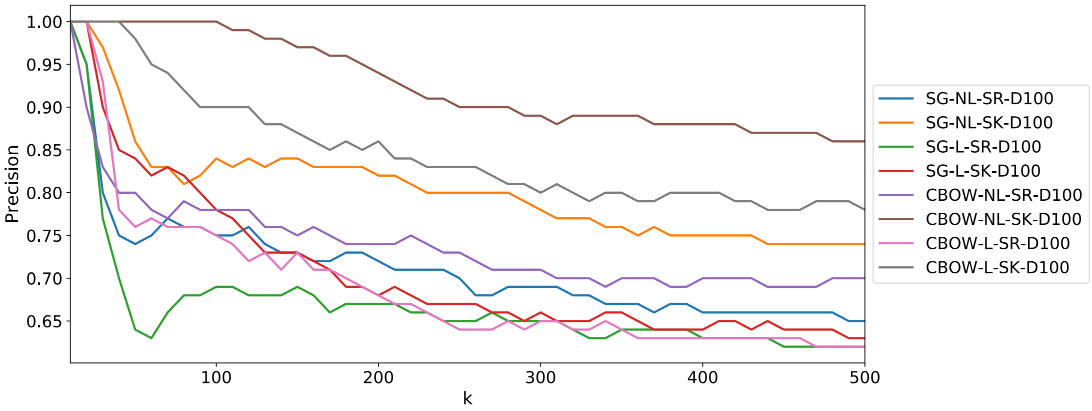

Figure 1: Precision@k results for models with $D = 100$ dimensions. Model labels are indicated as: Model (SG/CBOW), Lower / Not Lowercase (L/NL), Stop-word Remove/Keep (SR/SK).

Parameters	Values
Model	Skip-gram / CBOW
Dimension	25 / 100 / 400 / 800
Preprocessing	Lowercase / Stopwords

Table 1: Text preprocessing strategies and parameters evaluated.

along with initial close reading of the corpus:

health, disease, physic, physiology, pathology, therapeutics, remedies, medicine, physician, medical

For each embedding model, we extracted the top 20 terms that were most similar to the seed words used, and used these to build lexicons.

The evaluation of the lexicons involved using them as a basis for ranked document retrieval. We use a sample of 19,290 documents, each representing a labelled fixed-length excerpt from a text in the overall English corpus. Of these, 20% were labelled as medical texts. Texts were ranked according to the frequency of occurrence of terms from each generated lexicon. The quality of the lexicon was evaluated based on whether this ranking of documents surfaced texts related to medical topics. In this project, given the objective of enabling close reading of the retrieved texts within an exploratory interface, the precision of the returned results was of prime importance and evaluation was based on the level of precision relative to the top-k ranked texts (i.e. precision@k).

3 Findings and Analysis

Before considering document retrieval, we looked at the overall level of agreement between the lexicons generated by the models, by measuring their Jaccard set similarity for all 10 seed terms. We see a surprisingly low level of agreement between the lexicons – mean 0.31 and median 0.29.

Next, we measured the precision of the retrieval of medical documents for rankings of size $k \in [10, 500]$. The subset of results shown in Figure 1 reveal patterns indicating the importance of the choice of parameters and settings when generating word embeddings for literary and historical texts. Contrary to standard practice, not converting all text to lowercase and retaining stop-words resulted in better performance. This demonstrates that established standards for preprocessing modern texts may not produce the best results in a digital humanities context.

Error analysis, in the form of close reading, was conducted where strategies resulted in significantly lower precision (e.g. see SG-L-SR-D100 in Figure 1). These results appear to be due to the retrieval of country reports that, while they were not medical texts, contained a wealth of relevant information on medical care. This demonstrates how in a digital humanities project, error analysis can provide new information to prompt reformulation of the original research hypotheses.

4 Conclusion

This paper explores issues around selecting an appropriate strategy for using word embeddings to construct semantic lexicons for literary and historical texts. Established default strategies often emerge in response to requirements from different domains. However, this work shows the importance of evaluating the assumptions behind established strategies, and considering the specific requirements of individual digital humanities projects.

References

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2017. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *arXiv preprint arXiv:1707.01780* .

William August Guy. 1856. *Hooper's Physician's Vade Mecum*.

William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proc. EMNLP 2016*. NIH Public Access, volume 2016, page 595.

Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association of Computational Linguistics* 2(1):531–545.

Susan Leavy, Mark T Keane, and Emilie Pine. 2017. Mining the cultural memory of irish industrial schools using word embedding and text classification. In *Proc. Digital Humanities 2017*.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proc. ICLR 2013* pages 1–12.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

D Sculley and Bradley M Pasanek. 2008. Meaning and mining: the impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing* 23(4):409–424.

Gerhard Wohlgenannt, Ekaterina Chernyak, and Dmitry Ilvovsky. 2016. Extracting social networks from literary text with word embedding tools. In *Proc. Workshop on Language Technology Resources and Tools for Digital Humanities*. pages 18–25.