Domain specific text mining *almost* from scratch with Deep Learning

Linda Andersson
TU Wien,
Vienna, Austria
andersson@ifs.tuwien.ac.at

Abstract

Deep learning will help us to better design text mining applications, but perhaps not remove the computational linguistic design process associated with text mining applications (Manning, 2015). There has been extensive work on applying deep learning algorithms to different text mining applications such as information retrieval (IR) and information extraction (IE) and so far they have improved on classic IE and IR tasks. However, when deploying the algorithms on more advanced tasks, such as semantic role labelling, there is still some more work to be done (Collobert et al., 2011). In our research we compare and combine *traditional* natural language processing (NLP) techniques with distributional semantic models for domain specific retrieval.

Keywords: Patent Retrieval, Automatic Query Expansion (AQE), Distributional Semantics

1. Introduction

Within our research on the patent text domain, we demonstrate that distributional semantic methods such as word2vec can detect spelling errors (e.g. *calendering*, *calandering*), identify suffix or prefix (e.g. *calender*, *supercalender*) and lemmas of different word forms (e.g. *suppressing*, *suppress*, *suppressed*) and, for some specific words, it generates good synonyms (e.g. *underwear*, *underpants*, *undergarment*, *underclothes*). But for other words, such as *bus* and *cell*, the synonym suggestions are not useful as automatic query expansion (AQE) terms. We compare a word2vec model incorporated into the BM25 IR model (Rekabsaz et al., 2016; Rekabsaz et al., 2017) with a method which combines NLP techniques with word2vec representation in order to refine and correct domain specific lexical-semantic relations for automatic query expansion terms (Andersson et al., 2017). In (Andersson et al., 2017), they proposed a method to expand the cosine similarity computation for a unigram model of word2vec representation to include computation between concepts of arbitrary length. By combining NLP and the extended model, we can compute similarity between *bus slot card* and *ISA bus* versus *double-vehicle bus* and *automobile*.

2. Method

We experimented on the CLEF-IP 2013 test collection, which contains approximately 3M patent documents. For the query formulation (QF) method we re-used the best single word QF in (Andersson et al., 2016) and for the word2vec retrieval architecture we re-used methods presented by Rekabsaz et al. (2016). Rekabsaz et al. (2016) used word2vec to effectively change the content of the documents: for every query term with a similarity over a specific threshold (0.70 or 0.75) was added as an instance of the query term, weighted by its similarity score. For the NLP solution we used the retrieval architecture and the best QF method from (Andersson et al., 2016). Their best QF method (**NLP**) included a domain adapted NLP pipeline with additional machine learning for terminology extraction. In (Andersson et al., 2016) they assessed the performance both on the top 100 retrieved passages (paragraphs) and on document level for these top 100 retrieved passages. But since we only have the document representation for the **wd2v** method we only evaluate on document level. In this experiment we examine if we can improve on the **NLP** method by adding domain specific lexical-semantic relations (Andersson et al., 2014) for AQE or if we can substitute the entire process by switching to only using distributional semantic techniques. The ontology with the domain specific lexical-semantic relations was populated with lexico-syntactic patterns (Andersson et al., 2014) and a distributional semantic model was used to remove noisy terms and relations. The distributional semantic filter (**AQE SEM**) used a word2vec model, which was trained on patent data (300 dimensions). However, the generic word2vec techniques are limited in their deployment on patent text, since they are modelled upon unigram or a fixed length of n-grams (Mikolov et al., 2013). Meanwhile patent terms are a dynamic mixture of specific words and multi word terms (MWT) composed of common words of arbitrary length (Andersson et al., 2016; Judea et al., 2014). In order to expand the existing cosine computation to include computation between two MWTs of arbitrary length, we sum the similarity values of each combination and in order to avoid bias towards longer MWTs we divided the sum by the number of tokens:

$$Joined_{similarity} = \frac{\sum \left({}_{cos}(\overrightarrow{w_i}, \overrightarrow{w_n}) ... {}_{cos}(\overrightarrow{w_{i+1}}, \overrightarrow{w_n}) \right)}{N} \tag{1}$$

- ${}_{cos}(\overrightarrow{w_i},\overrightarrow{w_n})$ represents each word vector pair cosine similarity of a MWT; ${}_{n}$ the length of a MWT defined by its number of members.
- *N* is the number of words for a MWT.

3. Results & Conclusion

Table 1 shows the best performing methods in comparison with the state-of-the-art (Andersson et al., 2016) (**NLP**) and the best official participant run of the CLEF-IP 2013 passage retrieval task (Luo and Yang, 2013) (**Georgetown**). When applying word2vec representation within the BM25 model there is a decrease in performance for all metrics in

Table 1: CLEF-IP 2013 Passage retrieval task, evaluation on the top 100 passages on Document level

Run	Model	PRES	Recall	MAP
NLP AQE SEM (5)	LMJM	**0.558**	**0.649**	0.269
NLP	LMJM	0.544	0.631	**0.285**
v2wTreshold (0.75)	BM25	0.435	0.528	0.197
v2wTreshold (0.70)	BM25	0.435	0.528	0.196
Georgetown	BM25	0.433	0.540	0.191

comparison with the state-of-the-art method. Meanwhile, when applying **AQE SEM** filter, using 5 expansion terms we have a slight increase in performance for PRES (Magdy and Jones, 2010) and recall. As we can see, while results are visibly different, the relatively low number of topics (only 50 topics) in this track results in few clear cases of improvement. However, it gives us an indication how a deep learning algorithm performs in a linguistic challenging text domain.

4. Bibliographical References

Andersson, L., Lupu, M., Palotti, J., Piroi, F., Hanbury, A., and Rauber, A. (2014). Insight to hyponymy lexical relation extraction in the patent genre versus other text genres. In *Proc. of IPaMin KONVENS*.

Andersson, L., Lupu, M., Palotti, J., Hanbury, A., and Andreas, R. (2016). When is the time ripe for natural language processing for passage patent retrieval monitoring of vocabulary shifts over time. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM16.

Andersson, L., Rekabsaz, N., and Hanbury, A. (2017). Automatic query expansion for patent passage retrieval using paradigmatic and syntagmatic information. In *The first WiNLP workshop will be co-located with ACL 2017 in Vancouver*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

Judea, A., Schütze, H., and Brügmann, S. (2014). Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *COLING*, pages 290–300.

Luo, J. and Yang, H. (2013). Query formulation for prior art search-georgetown university at clef-ip 2013. In *Proc. of CLEF*.

Magdy, W. and Jones, G. J. (2010). Pres: A score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 611–618, New York, NY, USA. ACM.

Manning, C. D. (2015). Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Rekabsaz, N., Lupu, M., Hanbury, A., and Zuccon, G. (2016). Generalizing translation models in the probabilistic relevance framework. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 711–720.

Rekabsaz, N., Lupu, M., Baklanov, A., Hanbury, A., Dür, A., and Anderson, L. (2017). Volatility prediction using financial disclosures sentiments with word embedding-based ir models. In *Proc. of ACL*, Stroudsburg, PA, USA. Association for Computational Linguistics.