Improving automated content analysis with news-specific word embeddings for medium-resourced languages

Anne Kroon[⋆], Damian Trilling[⋆], Antske Fokkens[‡], Felicia Loecherbach[†], Judith Moeller[⋆], Mariken A.C.G. van der Velden[†], and Wouter van Atteveldt[†]

[⋆]Amsterdam School of Communication Research, University of Amsterdam
[†]Department of Communication Science, Vrije Universiteit Amsterdam
[‡]Computational Lexicology and Terminology Lab, Vrije Universiteit Amsterdam

Keywords: word embeddings, automated content analysis, news, small languages

Extended Abstract

The analysis of media content and political content is of crucial importance for many social-scientific disciplines. Manual content analysis gets more and more supplemented by or even replaced with automated methods. **Word embeddings** play an increasingly important role in various forms of automated text analyses. Word embeddings are typically high density (relatively) low-dimensional vector representations that provide meaning representations of words based on the context they occur in.

They are an important component in current state-of-the-art methods, such as information retrieval, sentiment analysis [4], machine translation and bias in text [1]. The performance of such downstream tasks, but also others such as topic classification or similarity detection [2], is highly dependent on the quality of the embeddings used. Consequently, high quality word embeddings can provide an important contribution for various methods often used in the field of social science. While there are many pre-trained models available for English, those who study content in other languages (such as Dutch) often have little to choose from.

In this contribution, we investigate whether it is worth the effort to train a custom model rather than relying on (limited) available pre-trained models. For the case of Dutch, few embedding models are available, and they are trained on ordinary human language from the World Wide Web. These models capture the specifics of news article data less well and are therefore likely to be less suitable to study and understand dynamics of this domain.

We aim to (1) develop and evaluate a high-quality embedding model; (2) assess performance in downstream tasks of interest to Communication Science (such as topic classification of newspaper data); (3) facilitate distribution and use of the model; (4) offer clear methodological recommendations for researchers interested using our Dutch embedding models.

We proceeded as follows. We retrieved $\approx$ 10 mln articles from Dutch news sources Telegraaf (print & online), NRC Handelsblad (print & online), Volkskrant (print & online), Algemeen Dabldad (print & online), Trouw (print & online), nu.nl , nos.nl, spanning two decades. We then compiled a corpus of all unique sentences, spanning 1.18B (1181701742) words and

77.1M (77151321) sentences. In particular, we trained a model using skip-grams with negative sampling, a window size of 5, and 300-dimensional word vectors. To do so, we used gensim's Word2Vec [3] package in Python.

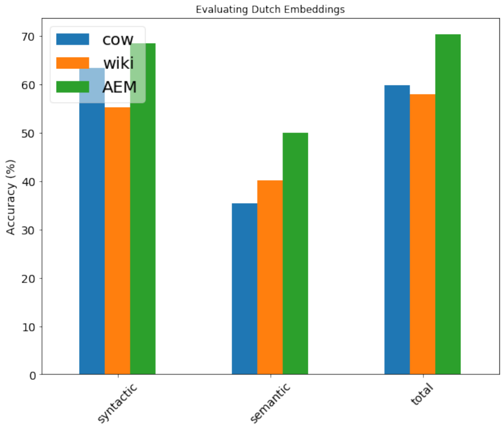

Figure 1: Internal evaluation.

We evaluate our Amsterdam Embedding Model ("AEM") by comparing it to two other publicly available embedding models (see Figure 1), in particular a FastText model trained on Wikipedia data ("wiki") and an embedding model trained on diverse .nl and .be sites [5, 6] ("cow"). We used a bench-marking task presented in [6]: A relation identification tasks analogous to evaluations on the English language which measures the quality of different kinds of word embeddings. Utilizing both semantic and syntactic accuracy indicators, we assess the model's performance in more than $>$ 5K relationship choice tasks.

Conclusions and future work

In this paper, we present a word embedding model trained on a Dutch news corpus, which outperforms existing models that were trained on the Dutch Wikipedia and various Dutch websites. This model can be used to improve the quality of automated content analysis in Dutch. In addition, our results suggest that training such models on news corpora are a feasible way of creating resources for languages with less resources than English. Our future work and work in progress includes the systematic evaluation of our model and other models in downstream classification tasks, such as news topic classification.

References

[1] Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science 356*, 6334 (2017), 183–186.

[2] Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. From word embeddings to document distances. In *Proceedings of The 32nd International Conference on Machine Learning* (2015), vol. 37, pp. 957–966.

[3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *NIPS*. Curran Associates, Inc., 2013, pp. 3111–3119.

[4] RUDKOWSKY, E., HASELMAYER, M., WASTIAN, M., JENNY, M., EMRICH, Š., AND SEDLMAIR, M. More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures 12*, 2-3 (2018), 140–157.

[5] SCHÄFER, R., AND BILDHAUER, F. Building large corpora from the web using a new efficient tool chain. In *Proceedings of LREC* (Istanbul, Turkey, 2012), Nicoletta Calzolari et al., Ed., ELRA, pp. 486–493.

[6] TULKENS, S., EMMERY, C., AND DAELEMANS, W. Evaluating unsupervised Dutch word embeddings as a linguistic resource. In *Proceedings of LREC* (2016), pp. 4130–4136.