Fankhauser, Peter ; Kupietz, Marc: Analyzing domain specific word embeddings for a large corpus of contemporary German. International Corpus Linguistics Conference, Cardiff, Wales, UK, July 22-26, 2019
DOI: https://doi.org/10.14618/ids-pub-9117

POSTPRINT

Analyzing domain specific word embeddings for a large corpus of contemporary German

Peter Fankhauser and Marc Kupietz
IDS-Mannheim, Germany

Introduction

Distributional models of word use constitute an indispensable tool in corpus based lexicological research for discovering paradigmatic relations and syntagmatic patterns (Belica et al. 2010). Recently, word embeddings (Mikolov et al. 2013) have revived the field by allowing to construct and analyze distributional models on very large corpora. This is accomplished by reducing the very high dimensionality of word co-occurrence contexts, the size of the vocabulary, to few dimensions, such as 100-200. However, word use and meaning can vary widely along dimensions such as domain, register, and time, and word embeddings tend to represent only the most prevalent meaning. In this paper we thus construct domain specific word embeddings to allow for systematically analyzing variations in word use. Moreover, we also demonstrate how to reconstruct domain specific co-occurrence contexts from the dense word embeddings.

Method and Corpus

To compute word embeddings, we employ the structured skip gram approach by Ling et al. (2015), using a one-hot encoding for words as input layer, a 200 dimensional hidden layer, and a positional one-hot encoding for the context words in a window of [-5,5] as the output layer. This approach takes word order into account, and thus also captures syntactic regularities of word usage.

As base corpus, we use the DeReKo-2017-II edition of the German Reference Corpus (Institut für Deutsche Sprache 2017; Kupietz et al 2010, 2018) containing 33 billion tokens. For the separation into domain slices, we use 11 of the top-level domain categories annotated in the metadata of DeReKo texts (Weiss 2005), discarding the domains *rest* and *unclassified*. The individual domains contain between 80 million and 9 billion tokens (see Figure 1).

For deriving domain specific word embeddings we adapt the approach of Dubossarsky et al. (2015), Fankhauser and Kupietz (2017), there used for diachronic corpora. We first compute embeddings for the base corpus and use them to initialize training of domain specific embeddings. Thereby the embeddings are comparable between individual domains and the base corpus.

One key question is, to which extent this two stage training regime can elicit domain specific word embeddings. To illustrate this, Figure 2 shows the distributions of the cosine distances (= 1 - cosine similarity) between domain specific word embeddings and their corresponding embeddings in the base corpus for selected domains, computed for the most frequent 30.000 types that occur at least 10 times in the given domain. Indeed, for quite a few types their domain specific embeddings end up fairly distant from their initial embeddings. The initial peak with distances near zero for the (smallest) domain *fiction* is due to types with rather few occurrences. In this

case, the domain specific training can apparently not override the initial embedding. When requiring at least 50 occurrences, this peak goes away. Mean (blue) and median (red) of the distributions lie between 0.08 for *politics* and 0.14 for *sports*, with the long tail reaching up to 0.8. The relative ranking of means corresponds well to other measures between domain specific word use, such as the Kullback Leibler divergence between the domain specific unigram language models (see below).

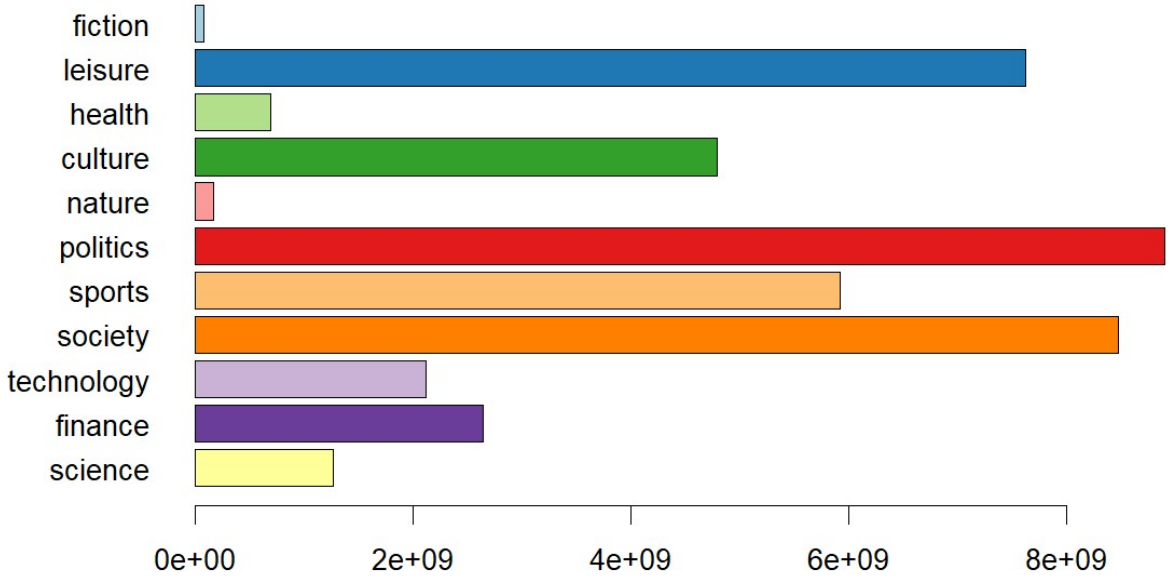

Figure 1. Domains and their token sizes

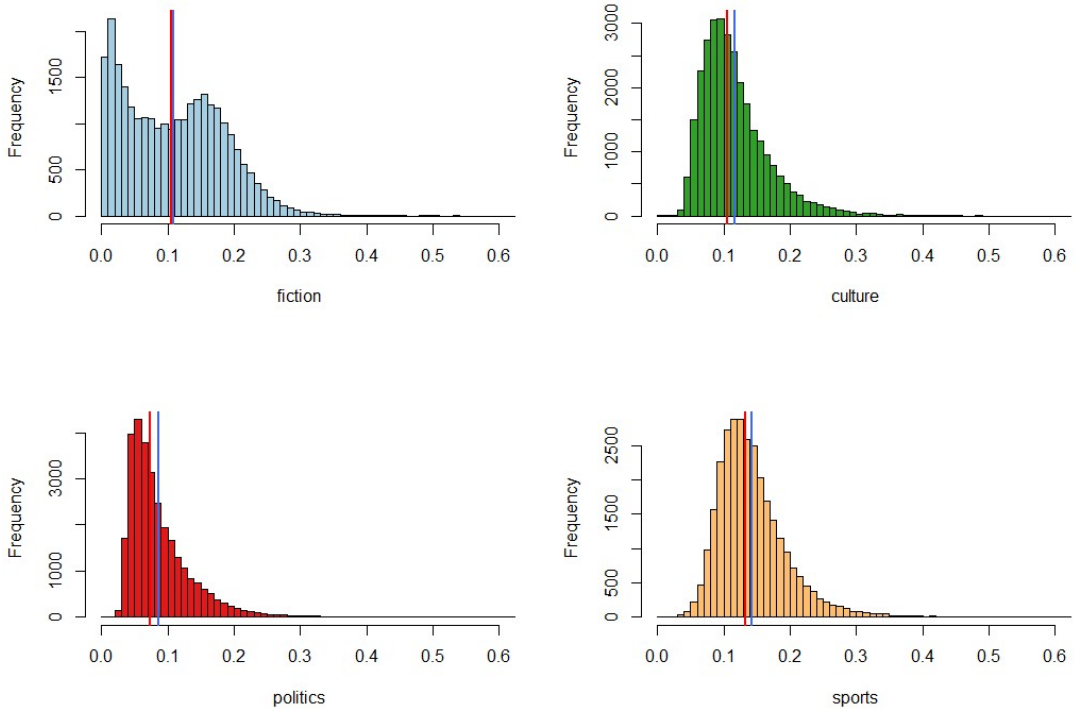

Figure 2. Distribution of Cosine Distances

Domain specific paradigmatic neighbourhoods

Investigating individual words reveals that the majority of large distances over 0.5 comes from proper nouns for persons, organizations, and locations. Another major kind of domain specific word use comes from word surface forms in different grammatical roles, such as adjective vs. participle (*"verspielt"*: *"playful"* vs. *"gambled away"* or *"verbissen"*: *"stubborn"* vs. *"bitten"*). However, also genuine lexicographic ambiguities can be identified, resulting in domain specific paradigmatic neighbourhoods. To visualize these we further reduce the 200 dimensions of word embeddings to two dimensions using t-Distributed Stochastic Neighbor Embedding (t-sne, Van der Maaten & Hinton 2008). Figure 3 shows the paradigmatic neighbourhoods of the word embeddings for *"Puppen"* in *DeReKo* vs. *science*.

Figure 3: Paradigmatic neighbours of *"Puppen"* (*"dolls"* / *"pupa"*) in DeReKo vs. science (green)

For a more general perspective beyond individual words, we also look at the overall distribution of domain specific paradigmatic neighbourhoods in the common semantic space. To this end, we determine for every word the domain for which it is most typical, where typicality of a word *w* is measured by its contribution to the Kullback-Leibler divergence between the domain and the base corpus (Fankhauser et al. 2014, Equation 1):

$$D(P\|Q) = \sum_{w} p(w) log_2 \left(\frac{p(w)}{q(w)} \right) \quad (1)$$

Table 1 gives the most typical words for selected domains, such as personal pronouns in *fiction*, animals in *nature*, and soccer related words in *sports*.

Using color for representing the most typical domain of a word, and the two dimensional t-sne representation of its embedding, we can visually correlate domain and semantics of words. Figure 4, left visualizes the fully zoomed out summary of the most frequent 30.000 words in the form of a bubble chart, with color of a bubble representing the domain along the color key in Figure 1, size representing its relative

frequency, and position representing its semantics. Even at this very abstract level, one can readily identify regions dominated by one or two domains. Figure 4, right shows two zoomed in regions. Region 1 is dominated by domain *nature* and comprises mainly words for animals, Region 2 is dominated by *culture* and comprises mainly kinds of cultural artefacts.

fiction	**culture**	**nature**	**politics**	**sports**
ich	und	Grad	die	gegen
daß	er	Tiere	der	den
er	Musik	Hund	dass	Trainer
sie	von	Hunde	SPD	Spiel
du	mit	Tierheim	für	Mannschaft
und	als	Tier	CDU	SV
mir	Film	Katzen	werden	FC
mich	Publikum	Temperaturen	nicht	Sieg
so	Band	Katze	sei	Saison
ihm	sie	Vierbeiner	des	Platz

Table 1: Most typical words for selected domains

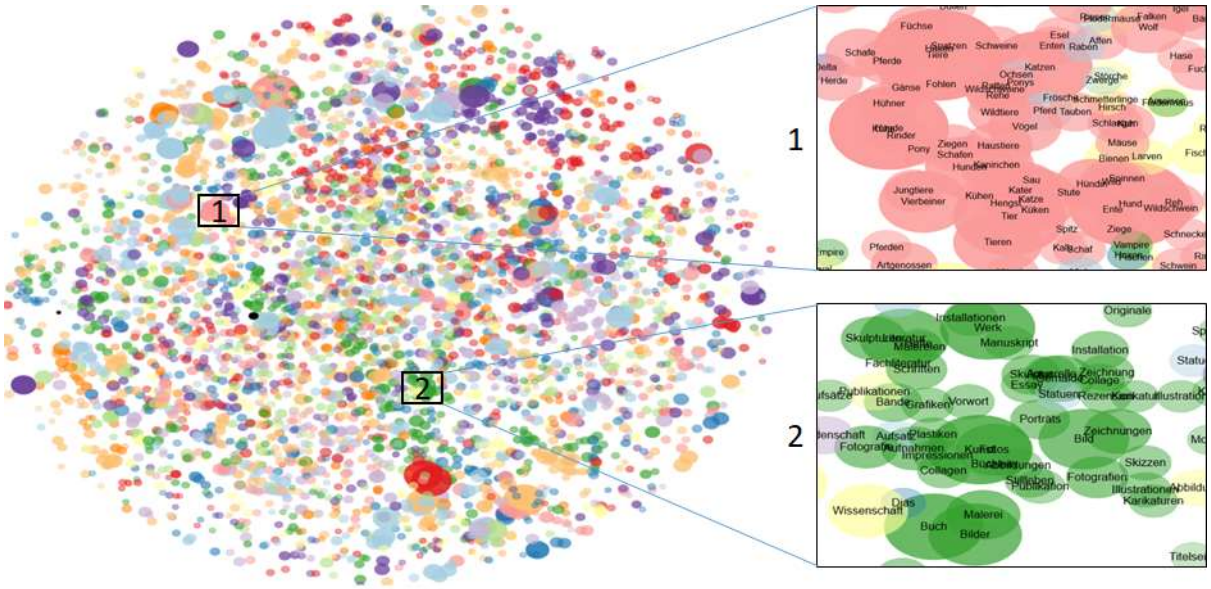

Figure 4. Visual Correlation of Domain and Semantics

In summary, domain specific word embeddings can not only be used to differentiate the domain specific paradigmatic neighbourhoods of individual words, but also to identify domain specific paradigmatic clusters.

Domain specific syntagmatic neighbourhoods

Word embeddings can only be interpreted by inspecting the paradigmatic neighbourhoods they induce. For a finer grained analysis, we need to reconstruct the syntagmatic context of words. In the structured skip-gram approach we use, this context is represented by means of the output layer, or, more specifically, by the connections between the hidden layer and the output layer (see also Levy & Goldberg

2014 and Kupietz et al. 2018). We can thus reconstruct the syntagmatic context of a target word by running one feed-forward cycle for its one-hot encoding in the input layer. The output layer then constitutes a joint positional one-hot representation of the target's context words, which approximates its underlying syntagmatic context. This one-hot representation can now be analyzed very much like traditional collocations: We use different scoring functions, such as average or maximum activation, to generalize over the different context positions and to select the most typical or salient syntagmatic neighbours.

For extracting and sorting the 10 most typical collocators of the example *"Puppen"* (Table 2) across different domains, we use total activation sum of a word averaged over all context positions at which it receives any activation (above a very low threshold).

dereko	**leisure**	**culture**	**sports**	**science**
Handpuppen	Plüschtiere	Handpuppen	tanzen	saugen
Schaukelpferde	Handpuppen	Fingerpuppen	auspackten	krabbeln
Puppenstuben	Puppenkleider	Stabpuppen	vorführen	Kokons
Puppenkleider	filzen	Marionetten	dressieren	Puppe
Teddies	Schaukelpferde	Holzpuppen	vorführten	bestäubt
basteln	Stofftieren	Masken	lassen	schlüpfenden
Teddys	Porzellanpuppen	Puppenstuben	dribbeln	abgesammelt
Stofftieren	Stofftiere	Schattenfiguren	verkleidet	verpuppt
Stofftiere	Teddys	REUTLINGEN	hochheben	fressen
töpfern	Puppenstuben	Spejbl	Salti	ähneln

Table 2: Most typical syntagmatic neighbours of *"Puppen"* (*"dolls"* | *"puppets"* | *"pupa"*) across selected domains

The analysis of the typical syntagmatic neighbours of *"Puppen"* shows close synonyms of the *"dolls"* reading for *DeReKo* and for the *leisure* domain (due to enumerations). In contrast, the *culture* domain shows mainly close synonyms for a *"puppets"* reading, whereas the *science* domain shows different types of collocations for the botanic *"pupa"* reading. Finally, the *sports* domain focuses on collocations of more figurative usages of *"Puppen"* (like in "die Puppen tanzen lassen", roughly: "to paint the town red").

Conclusions

We have presented an approach to construct domain specific word embeddings. On this basis we have analyzed the distribution of distances between word embeddings for selected domains, and shown that the approach indeed can differentiate between domain specific word usage. Inspecting the paradigmatic neighbourhood of words induced by their embeddings, we have given examples of words with multiple meanings, and shown how individual domains populate the semantic space in the form of closely knit clusters of paradigmatically related words. Moreover, we have shown how word embeddings together with their output layer can be used to also analyze the syntagmatic neighbourhood of words for a finer grained differentiation of multiple

meanings. All examples and visualizations have been generated from an interactive visualization which is available at:
http://corpora.ids-mannheim.de/openlab/sliceviz/description.html.

References

Belica, C., Keibel, H., Kupietz, M., Perkuhn, R. (2010): An empiricist's view of the ontology of lexical-semantic relations. In Storjohann, P. (ed.): Lexical-Semantic Relations. Theoretical and practical perspectives. Amsterdam: Benjamins, 115-144.

Dubossarsky, H., Tsvetkov, Y., Dyer, C., Grossman, E. (2015). A bottom up approach to category mapping and meaning change. In Pirrelli, Marzi & Ferro (eds.), Word Structure and Word Usage. Proceedings of the NetWordS Final Conference.

Fankhauser, P., Knappen, J., Teich, E (2014). Exploring and Visualizing Variation in Language Resources. In Proceedings of the ninth international conference on language resources and evaluation (LREC '14)

Fankhauser, P., Kupietz, M. (2017). Visualizing Language Change in a Corpus of Contemporary German. In Proceedings of the 9th International Corpus Linguistics Conference, University of Birmingham, Tuesday 25–Friday 28 July 2017

Institut für Deutsche Sprache (2017): *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2017-I* (Released 2017-03-08). Mannheim: Institut für Deutsche Sprache. PID: 10932/00-0373-23CD-C58F-FF01-3. http://www.dereko.de/

Kupietz, M., Belica, C., Keibel, H., Witt, A. (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta: ELRA, 1848-1854.

Kupietz, M., Lüngen, H., Kamocki, P., Witt, A. (2018). The German Reference Corpus DeReKo: New Developments – New Opportunities. In: Calzolari, N. et al (eds): Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki: ELRA, 4353-4360

Levy, O., Goldberg, Y. (2014): Neural Word Embedding as Implicit Matrix Factorization. NIPS.

Ling, W., Dyer, C., Black, A., & Trancoso, I. (2015). Two/too simple adaptations of word2vec for syntax problems. In Proc. of NAACL.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of NIPS 2013, 3111–3119.

Van der Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE. In Journal of Machine Learning Research 1, 1-48.

Weiß, C. (2005): Die thematische Erschließung von Sprachkorpora. In: OPAL - Online publizierte Arbeiten zur Linguistik 1/2005. Mannheim: Institut für Deutsche Sprache, 2005.