Count-Based and Predictive Language Models for Exploring DeReKo

Peter Fankhauser, Marc Kupietz
Leibniz Institute for the German Language
Mannheim, Germany
{fankhauser|kupietz}@ids-mannheim.de

Abstract

We present the use of count-based and predictive language models for exploring language use in the German Reference Corpus DeReKo. For collocation analysis along the syntagmatic axis we employ traditional association measures based on co-occurrence counts as well as predictive association measures derived from the output weights of skipgram word embeddings. For inspecting the semantic neighbourhood of words along the paradigmatic axis we visualize the high dimensional word embeddings in two dimensions using t-stochastic neighbourhood embeddings. Together, these visualizations provide a complementary, explorative approach to analysing very large corpora in addition to corpus querying. Moreover, we discuss count-based and predictive models w.r.t. scalability and maintainability in very large corpora.

Keywords: language models, word embeddings, collocation analysis

1. Introduction

Distributional semantics is concerned with analysing language use based on the distributional properties of words derived from large corpora. In this paper we describe DeReKoVecs[1] (Fankhauser and Kupietz, 2017), a visualization of distributional word properties derived from the German Reference Corpus DeReKo[2] (Kupietz et al., 2010) comprising more than 53 billion tokens of written contemporary German.

DeReKoVecs represents the syntagmatic context of words in a window of five words to the left and to the right $w_{-5} \dots w_{-1}\, w\, w_1 \dots w_5$ as vectors. These vectors are either count-based or predictive.

The count-based models are computed by various association measures based on (co-occurrence) frequencies in the corpus; for an overview see e.g. Evert (2008).

The predictive models are trained using structured skipgrams (Ling et al., 2015), an extension of word2vec (Mikolov et al., 2013) that represents the individual positions in the syntagmatic context of a word separately, rather than lumping them together into a bag of words.

Figures 1 and 2 compare count-based and predictive models for a word w in its left/right syntagmatic context with collocates $w_{-2}\, w_{-1}\, \_\, w_1\, w_2$.

The count-based model represents each pair $w_i\, w$ individually by some association measure o_i. With a vocabulary size of v (the number of different words, aka types) this leads to a very high dimensional model with order $O(v^2)$ parameters, where each word is represented by a sparse vector of size $4 * v$.

In contrast, the predictive model introduces a hidden layer h of size d. d is typically in the range of 50 to 300 and thus much smaller than v, which in the case of DeReKo ranges in the millions. Each word can thereby be represented by a much smaller vector of size d, also called its word embedding. Importantly, estimates of the association strength between w and its left and right collocates can still be gained via its output activations[3].

Both models support the analysis of word use along the paradigmatic and the syntagmatic axis. Paradigmatically related words, such as synonyms or (co-)hyponyms, which occur in similar syntagmatic contexts, can be identified by determining the similarity (usually cosine similarity) between their vectors, which are, by construc-

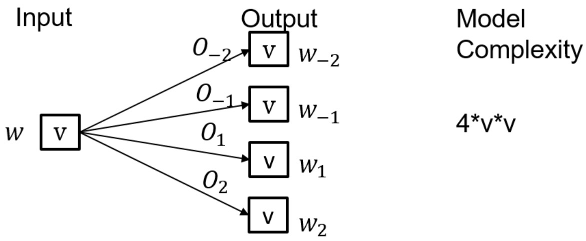

Figure 1: Count-based Model

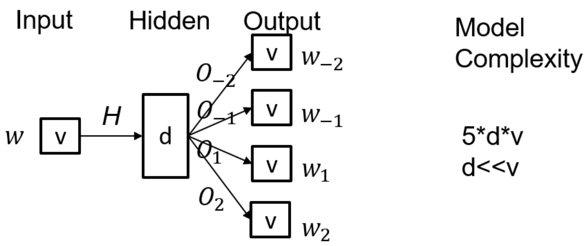

Figure 2: Predictive Model

[1] http://corpora.ids-mannheim.de/openlab/derekovecs

[2] https://www1.ids-mannheim.de/kl/projekte/korpora/

[3] More specifically, the output activations approximate the shifted pointwise mutual information. $SPMI(w, w_i) = log(\frac{p(w,w_i)}{p(w)p(w_i)}) - log(k)$, with k the number of negative samples used during training (see Levy and Goldberg 2014). Pointwise mutual information is one of the count-based collocation measures in DeReKoVecs.

Proceedings of the CMLC-10 Workshop (@LREC 2022), pages 27–31
Marseille, 20 June 2022

Kuh	German	English
Count	Kalles **heilige blöde Blinde Bunte** lila Rosmarie **dumme** Yvonne **Eis**	Kalle's **holy silly blind colorful** purple Rosemary **stupid** Yvonne **ice**
Pred	ausgebüxte geschlachtete entlaufene geklonte trächtige geschlachteten weidende verwesende Kalles tote	escaped slaughtered runaway cloned pregnant slaughtered grazing decaying Kalle's dead

Table 1: Count-based and predictive collocates for 'Kuh' ('cow')

tion, a representation of their syntagmatic contexts. Syntagmatically related words, which occur close to each other more often than expected, are represented by their count-based or computed association strength.
Count-based models and predictive models complement each other. Count-based models excel at representing all actually occurring, possibly polysemous usages, but they just memorize and do not generalize to other possible usages. In particular, they can fail to adequately represent low frequency words and collocations for which there simply do not exist enough examples. Predictive models generalize by means of dimensionality reduction in the hidden layer and thus can also predict unseen but meaningful usages, but they typically only represent the dominant, usually literal usage [4].
In the following we illustrate the interplay between count-based and predictive models along the syntagmatic and the paradigmatic axis by way of example.

2. Syntagmatic Analysis

Tables 1, 2 and 3 exemplify the interplay between count-based and predictive collocations[5].
Among the top 10 count-based collocates of 'Kuh' (cow), there are 6 collocates (in bold) stemming from idiomatic use, for example, 'die Kuh vom Eis kriegen' literally for 'getting the cow from the ice' meaning 'working out a situation'. In contrast, the predictive collocates all pertain to the literal meaning of cow as a (domestic) animal; e.g., 'Eis' does not occur among the top 400 predictive collocates.

[4]This focus on the dominant usage may be one of the main reasons for the relative success of predictive models as opposed to count-based models for lexical semantics tasks observed in (Baroni et al., 2014), as these tasks tend to focus on dominant semantics.

[5]We employ a variety of measures for the association strength between collocates. Here we only use the default measures: LogDice for count-based and the sum of output weights for the given word w normalized by the total weights for all words w_i. Both are restricted to those words w_i which maximize the measure.

Versuch	German	English
Count	unternommen gescheitert Beim zweiten gescheiterten wert dritten gestartet unternehmen scheiterte	made failed in second failed worth third started make failed
Pred	untauglicher vergeblicher missglückter unternommene krampfhaften fehlgeschlagener (…)	unsuitable futile failed made convulsive failed failed desperate unsuitable desperate

Table 2: Count-based and predictive collocates for 'Versuch' ('attempt')

Absatz	German	English
Count	**reißenden** Paragraf Paragraph **fanden** Berichtigung Satz Zeile **Reißenden** Grundgesetzes Aktualisierung	**soaring** paragraph **found** correction sentence line **soaring** constitution update
Pred	**reißenden reissenden rückläufigem** Unsinniger **Sinkender** bequellt **stagnierendem** unbelegten **reißend sinkendem**	**soaring declining** meaningless **decreasing** quoted/sourced **stagnant** unsubstantiated **soaring decreasing**

Table 3: Count-based and predictive collocates for 'Absatz' ('paragraph' vs. 'sales')

The count-based and predictive collocates of 'Versuch' ('attempt'), on the other hand, show no such difference. Both refer to the literal meaning of 'Versuch'. However, also here we can observe a bias of the predictive collocates towards a dominant usage as in 'failed attempts'.

Finally, the count-based and predictive collocates of 'Absatz' in Table 3 both comprise two usages/meanings: 'paragraph' and 'sales' (in bold). However, in particular the top count-based collocates for 'Absatz' as in 'sales' stem all from the fixed phrase 'reißenden Absatz finden' (literally: 'find soaring sales', roughly: 'sell like hotcakes'), whereas the predictive collocates cover a broader range of usages.

In summary, count-based collocates tend to come from fixed, possibly idiomatic phrases, whereas predictive collocates generalize to a broader range of words pertaining to a dominant meaning. An application of this discrepancy to detecting German idioms is described in Amin et al. (2021a; 2021b).

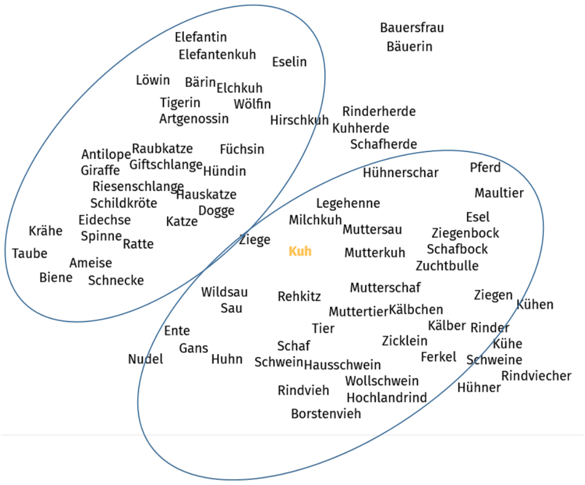

Figure 3: Paradigmatic neighbourhood of ‘Kuh’

Figure 4: Paradigmatic neighbourhood of ‘Versuch’

3. Paradigmatic Analysis

Looking at the paradigmatic axis for words with a similar usage context corroborates the syntagmatic analysis. Currently, we only provide for paradigmatic analysis on the basis of the predictive models but not based count-based models. For visualization we use t-stochastic neighbour embedding (t-sne, Van der Maaten and Hinton (2008)). T-sne maps the cosine distance between the high (200) dimensional word representations to two dimensions, such that small, local distances are preserved well, whereas global distances are not[6].

Figure 3 depicts the paradigmatic neighbourhood of ‘Kuh’ (‘cow’). We can observe two main clusters, both referring to the literal meaning[7]. The top left cluster comprises wild animals, largely but not exclusively mammals, and the bottom right cluster comprises farm animals. The idiomatic use of ‘Kuh’ is not reflected[8].

The paradigmatic neighbourhood of ‘Versuch’ (‘attempt’, Figure 4) can be roughly divided into three clusters. ‘Versuch’ as a mental process (top left), ‘Versuch’ as a trick (top right), and as an action, usually expressed via a composite word (bottom).

Both ‘Kuh’ and ‘Versuch’ arguably only depict one broad meaning clustered into fine but nonetheless meaningful nuances. In contrast, the paradigmatic neighbourhood of ‘Absatz’ shown in Figure 5 gets clearly separated into ‘paragraph’ (left) and ‘sales’ (right). These two individual broad clusters can again be divided into fine grained subclusters (e.g. ‘article’, ‘sentence’, ‘section’ for ‘paragraph’), but the big divide between ‘paragraph’ and ‘sales’ along the syntagmatic axis for both, the count-based and the predictive model, also shows along the paradigmatic axis.

4. Performance & Maintainability

An important motivation for us to experiment with word embedding models was the expectation that, thanks to efficient dimension reduction, they would be more performant to compute and more efficient to analyse in terms of paradigmatic neighbourhoods than the count-based models used so far in the context of the CCDB platform (Keibel and Belica, 2007).[9]

For the latter, the necessary precalculation of paradigmatic distances was considered to be so computationally expensive that it was hardly maintainable and the last calculation was carried out on the basis of DeReKo-2006-I, so that distributional analyses of the very current language use, based on DeReKo, was not possible for a long time.

We cannot yet draw a final conclusion regarding the performance comparisons, since we have not yet implemented paradigmatic analyses based on the count-based models. However, the computation time of the word embedding network for DeReKo-2022-I (53G tokens) is with 10 days roughly equivalent to the creation of a corresponding co-occurrence database,[10] each with 10 context words.[11] The disk space requirement is slightly larger with 61,2 GB vs. 45 GB in the case of the word embeddings.

As far as the runtime behaviour is concerned, it should be noted that for the calculation of the syntagmatic neighbours, the entire word embedding network is kept virtually in memory via memory mapping, so that if many

[6] Our visualization also provides for self organizing maps (SOM) (Kohonen, 1982), which position paradigmatic neighbourhoods on a grid of 6x6 squares.

[7] The ellipses are manually superimposed for the purpose of illustration.

[8] Incidently it is also not reflected in the count-based paradigmatic neighbourhood, not shown here.

[9] http://corpora.ids-mannheim.de/ccdb/

[10] based on RocksDB (Dong et al., 2021)

[11] on a Supermicro Intel(R) Xeon(R) Gold 6148 CPU Linux server with 80 cores @ 2.4 GHz and 756 GB RAM

Figure 5: Paradigmatic neighbourhood of 'Absatz'

instances are required, the RAM requirement can become a bottleneck.

All in all, both the calculation and the runtime behaviour are in a range that allows an annual update and the continuous operation of up to five instances, in our case. The approach is also quite scalable. The calculation of the predictive models can be accelerated by using more processor cores and building the count-based model with faster disks. The integrity of the programmes is ensured by CI workflows with an increasing number of tests, maintainability by a small number of dependencies, and easy deployment by Dockerization. Only the extension is somewhat challenging, as the code is mainly written in C, C++ and Perl.[12]

5. Availability

All tools that have been used in this paper to compute and analyse the models and to visualize the results are published under the Apache License 2.0 and available open-source on our Gerrit code-review site.[13]

We are happy to share all count-based and predictive models with interested colleagues under the Text and Data Mining exception (§ 60d German Copyright Act) (see also Kamocki et al. 2018).

6. Conclusions

We have described the implementation and use of count-based and predictive models for syntagmatic and paradigmatic analysis of language use in the German Reference Corpus DeReKo. Currently, we work on two main lines of extending the presented approach: (1) To allow a more principled comparison between count-based and predictive association measures, we plan to map the output weights to actual co-occurrence predictions. (2) To be able to contrast language use in different contexts, such as register or time, we experiment with several approaches to train context-dependent word embeddings. Finally, we also plan to apply the presented approach to other corpora.

7. Acknowledgements

We thank our anonymous reviewers for helpful comments on earlier drafts of the manuscript.

8. Bibliographical References

Amin, M., Fankhauser, P., Kupietz, M., and Schneider, R. (2021a). Data-driven identification of idioms in song lyrics. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 13–22, Stroudsburg, PA. Association for Computational Linguistics.

Amin, M., Fankhauser, P., Kupietz, M., and Schneider, R. (2021b). Shallow context analysis for german idiom detection. In *Proceedings of the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021*.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

Diewald, N., Margaretha, E., and Kupietz, M. (2021). Lessons learned in quality management for online research software tools in linguistics. In Harald Lüngen,

[12] see Diewald et al. (2021) for the relevance of such aspects for linguistic research (tools)

[13] https://korap.ids-mannheim.de/gerrit/plugins/gitiles/ids-kl/dereko2vec
https://korap.ids-mannheim.de/gerrit/plugins/gitiles/ids-kl/derekovecs

et al., editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event)*, pages 20 – 26, Mannheim. Leibniz-Institut für Deutsche Sprache.

Dong, S., Kryczka, A., Jin, Y., and Stumm, M. (2021). Rocksdb: Evolution of development priorities in a key-value store serving large-scale applications. *ACM Trans. Storage*, 17(4), oct.

Evert, S. (2008). Corpora and collocations. In Anke Lüdeling et al., editors, *Corpus Linguistics. An International Handbook*, chapter 58, pages 1212–1248. Mouton de Gruyter, Berlin, Germany.

Fankhauser, P. and Kupietz, M. (2017). Visualizing language change in a corpus of contemporary german. In *Corpus Linguistics Conference*, Birmingham, United Kingdom.

Kamocki, P., Ketzan, E., Wildgans, J., and Witt, A. (2018). New exceptions for Text and Data Mining and their possible impact on the CLARIN infrastructure. In Inguna Skadina et al., editors, *CLARIN Annual Conference 2018, Proceedings. 8-10 October 2018, Pisa, Italy*, pages 39 – 42, Utrecht. CLARIN.

Keibel, H. and Belica, C. (2007). CCDB: A corpus-linguistic research and development workbench. In *Proceedings of the 4th Corpus Linguistics Conference (CL 2007)*, Birmingham.

Kohonen, T. (1982). Self–organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In Nicoletta Calzolari, et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1848–1854, Valletta/Paris. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*.

Ling, W., Dyer, C., Black, A. W., and Trancoso, I. (2015). Two/too simple adaptations of Word2Vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304, Denver, Colorado, May–June. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Van der Maaten, L. and Hinton, G. E. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, November.