Healthcare of the Future 2022
T. Bürkle et al. (Eds.)

doi:10.3233/SHTI220314

Evaluation of Domain-Specific Word Vectors for Biomedical Word Sense Disambiguation

Dennis TODDENROTH [1]
Chair of Medical Informatics, University Erlangen-Nuremberg, Germany

Abstract. Among medical applications of natural language processing (NLP), word sense disambiguation (WSD) estimates alternative meanings from text around homonyms. Recently developed NLP methods include word vectors that combine easy computability with nuanced semantic representations. Here we explore the utility of simple linear WSD classifiers based on aggregating word vectors from a modern biomedical NLP library in homonym contexts. We evaluated eight WSD tasks that consider literature abstracts as textual contexts. Discriminative performance was measured in held-out annotations as the median area under sensitivity-specificity curves (AUC) across tasks and 200 bootstrap repetitions. We find that classifiers trained on domain-specific vectors outperformed those from a general language model by 4.0 percentage points, and that a preprocessing step of filtering stopwords and punctuation marks enhanced discrimination by another 0.7 points. The best models achieved a median AUC of 0.992 (interquartile range 0.975 - 0.998). These improvements suggest that more advanced WSD methods might also benefit from leveraging domain-specific vectors derived from large biomedical corpora.

Keywords. Word sense disambiguation, word vectors, linear classifiers.

1. Introduction

The increasing electronic availability of various biomedical texts presents novel opportunities to reuse such information resources with methods from computational linguistics, or natural language processing (NLP). Medical or scientific NLP applications may for example verify the completeness of clinical documentation [1], or compute and evaluate concise summaries of clinical trial descriptions [2].

Relevant NLP subtasks include word sense disambiguation (WSD), which aims to estimate intended homonym meanings from surrounding text. Typically, human readers can intuitively infer from context whether a mentioned word such as *cortex* refers to a part of the brain or to a substructure of the adrenal glands. Due to the variety of enclosing syntax and semantics, however, implementing such a function in an NLP algorithm can be rather difficult.

Recently developed NLP methods include word vectors or embeddings that derive high-dimensional numeric representations from processing local co-occurrence patterns in large corpora. Since related words tend to be used in similar contexts, such word vectors reflect semantic similarity as proximity in vector space. Word vectors thus encode meaning in a form that combines computability with nuanced semantics, which

[1] Corresponding Author; E-mail: dennis.toddenroth@fau.de.

may exceed the capabilities of discrete vocabularies and explicitly defined NLP algorithms.

The effectiveness of word vectors depends on the size of the training corpora, as well as on their representativeness for subsequent textual inputs. The authors of *scispaCy* [3], a modern NLP library that includes word vectors trained in vast biomedical corpora, therefore propose that their package should be particularly suitable for processing domain-specific texts. Below we explore the utility of *scispaCy* word vectors for biomedical WSD by training and evaluating a set of simple linear classifiers.

2. Methods and Data

To explore the utility of domain-specific word vectors for WSD, we consider homonyms that are prevalent in Pubmed abstracts. Pubmed queries that leverage the controlled vocabulary used to index the underlying literature database (Medical Subject Headings, or MeSH) can produce highly selective results when search terms are represented in MeSH terms.

To disambiguate words with technical meanings that are increasingly used in publications from the field of digital medicine, we manually labeled 1,493 abstracts. These annotations resolve words such as *Matlab*, which is not part of MeSH, and which can either denote a programming language or a region in Bangladesh. Other labels differentiate whether *Java* refers to a programming language or to an island in Indonesia, whether *Python* denotes a serpent, and whether *R* is used in the sense of a statistical correlation. While alternative meanings may occur with different frequencies in the biomedical literature, annotations were assigned so that the frequencies of alternative meanings are rather similar (see legend in figure 1); this procedure aimed to optimize the efficiency of the analysis by maximizing the number of pairs of abstracts from which linguistic differences between alternative contexts can be derived. Another part of the analysis involved a subset of four medical homonyms that had been semi-automatically annotated in 923 abstracts from the MSH WSD data set in order to support the development and assessment of such algorithms [4].

Word vectors are originally associated with single tokens, but document-specific vectors can be computed by aggregating vectors weighted by the frequency of mentioned words. As a consequence of their high dimensionality, these aggregate vectors may encode the “average meaning” of a document or context remarkably well, even when the information contained in word sequences is entirely discarded. The left side of figure 1 shows the first two dimensions of a principle component analysis of abstract-specific vectors for the first one of the described data sets, computed via the R function *prcomp()*, colored according to the respective WSD labels.

Even if this image obscures substantial information from the other dimensions with lower variance, we can see some label disaggregation, which can in turn be used to estimate annotations. The arrows depicted in figure 1 represent pairwise differences between projected average locations or gravity centers, so that the scalar product of each (un-projected) directional vector and those of separate samples optimizes the association with the respective labels.

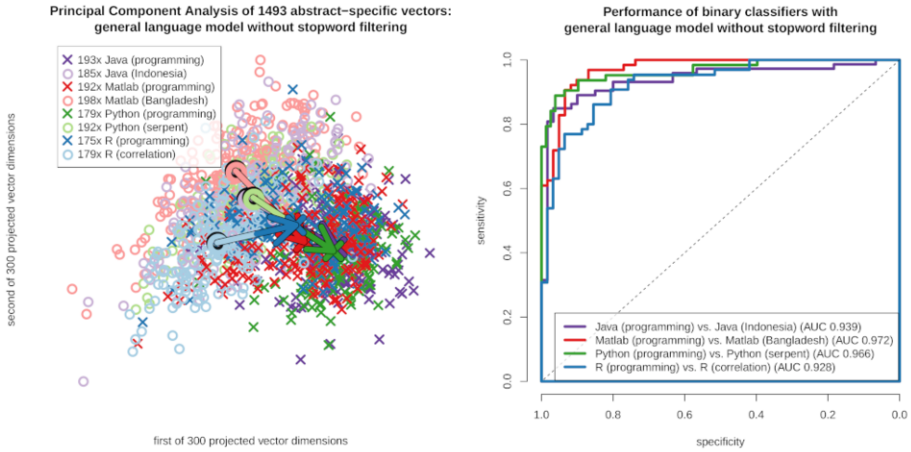

Figure 1. Principle components of word vectors for 1493 abstracts, colored according to manually defined annotations; arrows highlight pairwise differences between projected label-specific average locations (left). Receiver Operating Characteristic curves demonstrate how well a linear classifiers trained in a bootstrap subsample reproduces disjunctive held-out labels (right).

Natural language contains frequent words such as articles and conjunctions that are important for comprehending coherent texts, while their sheer presence of absence is relatively unspecific for alternative homonym contexts. Since the described aggregation already ignores word sequences, punctuation marks and so-called stopwords may be seen as noise in terms of the "average meaning" in vector space. Therefore, we evaluated the effect of removing stopwords on WSD performance.

From these data sets with 1,493 and 923 pairs of vectors and labels, partitions with disjunctive training and evaluation subsamples were repeatedly selected via bootstrapping. This procedure trained classifiers on rows that were randomly sampled with replacement, while disjunctive hold-out samples served to evaluate the respective classifier (200 repetitions). WSD performance was then measured as the area under Receiver Operating Characteristic curves that summarize the achievable constellations of sensitivity and specificity (ROC-AUC as in figure 1).

Word vectors were calculated under Python 3.8 with *spaCy* 3.0 and its general *en_core_web_md* language model (vectors with 300 dimensions) as well as *scispaCy*'s domain-specific *en_core_sci_lg* model in version 0.4 (200d). Classifier training and evaluation was implemented in R 3.6. Annotation data, script sources, and computed vectors are available at https://www.github.com/dtoddenroth/embeddingswsd/.

3. Results

For eight homonyms, four algorithmic configurations and 200 bootstrap repetitions, WSD effectiveness as measured by the median ROC-AUC in a total of 6,400 models was 0.976, with an overall range from 0.748 to 1.000 and an interquartile range from 0.945 to 0.991. Figure 2 summarizes the respective ROC-AUC distributions for WSD tasks and different algorithmic configurations. Note that the displayed vertical scale in figure 2 is restricted to ROC-AUC ranges between 0.8 and 1.0.

When aggregating model performance across tasks and bootstrap repetitions, we find that classifiers trained on domain-specific vectors outperformed those from a general language model by 4.0 percentage points (percentages of the unit square under a ROC curve), while a preprocessing step of filtering stopwords and punctuation marks enhanced discrimination by another 0.7 points. After pooling AUC values from WSD tasks and bootstrap repetitions in order to compare algorithmic configurations, the best constellation that combined stopword filtering with the specialized biomedical language model achieved a near-optimal median AUC of 0.992 (interquartile range 0.975 to 0.998).

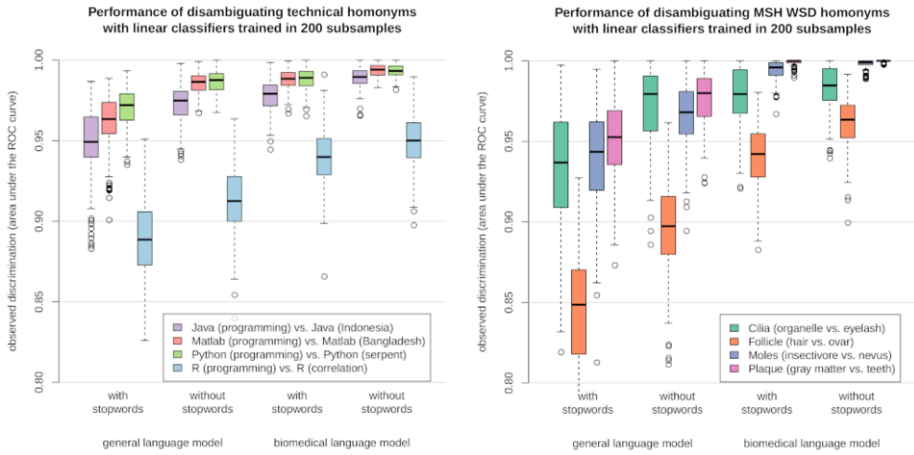

Figure 2. Tukey's boxplots visualize the variance of classifier performance (ROC-AUC) in 200 bootstrap partitions for disambiguating four technical homonyms (left) and four homonyms from the MSH WSD data set (right) for different vector sets and algorithmic configurations.

4. Discussion

Our findings indicate that linear classifiers based on domain-specific vectors outperformed those from the general language model. A preprocessing step of filtering stopwords and punctuation marks also improved model discrimination. The difficulty of the tested disambiguation tasks seemed to vary, and the most effective classifiers achieved near-optimal performance on the easier tasks.

The described method may seem naïve in the sense that it ignores all information contained in the particular word sequences around homonyms. The corresponding computational simplicity based on linear algebra, on the other hand, could facilitate implementing the approach for wider usage. When meaningful vector representations of documents can be stored in a relational database or in processing units that are optimized for parallel matrix multiplication, it seems feasible that vector-based document filtering and ordering can be realized with favorable flexibility and scalability.

Suppose that a researcher prepares a manuscript and wants to round out her set of cited references with additional relevant sources. While explicit search terms might of course be useful for finding candidate publications, homonyms among feasible criteria such as *follicle* might retrieve heterogeneous suggestions from dermatology as well as from gynecology, while her research may be concerned with only one of these topics. A

word vector derived from her manuscript or from the abstracts of cited references might then discriminate relevance more precisely than any constellation of search terms. Due to their nuanced high-dimensional representation, task-specific word vectors could even outperform explicit queries that do not include any homonyms. Since explicit MeSH labels are characteristically available for a subset of publications in the entire Pubmed literature database, future research might study the comparative or complementary value of vector-based predictions for WSD tasks, or could explore how well vector-based NLP models can impute MeSH labels for un-indexed articles.

Beyond the biomedical literature, querying textual patient records or clinical trial descriptions might also constitute promising use cases for classifiers based on word vectors. Medical informatics researchers have for some time developed and evaluated computerized trial recruitment systems that aim to support patient enrollment in ongoing medical studies, including with NLP methods [5] and machine learning models [6]. While we cannot expect word vectors to adequately capture all details such as intricate temporal criteria [7], part of the difficulty of matching patients to suitable trials may be attributable to ambiguous phrases in textual descriptions. If an eligibility criterion for example refers to a condition such as *plaque*, a word vector from its context could automatically suggest fitting neurological or dental patient records. Future studies might therefore explore whether the WSD effectiveness that we observed in scientific texts can be reproduced in individual patient-level clinical documents.

Previous research has deployed vector calculus for biomedical WSD [8], including in conjunction with recurrent convolutional neural networks based on self-trained embeddings [9]. The advanced performance that we observed when using simple linear classifiers based on domain-specific vectors suggests that future research might investigate whether such refined sequence-aware WSD models will also benefit from using vectors that had been pre-trained on large biomedical corpora.

References

[1] Schulz S, Seddig T, Hanser S, Zaiss A, Daumke P. Checking coding completeness by mining discharge summaries. Stud Health Technol Inform. 2011;169:594-8.

[2] Gulden C, Kirchner M, Schüttler C, Hinderer M, Kampf M, Prokosch HU, Toddenroth D. Extractive summarization of clinical trial descriptions. Int J Med Inform. 2019 Sep;129:114-121. doi: 10.1016/j.ijmedinf.2019.05.019.

[3] Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Proceedings of the 18th BioNLP Workshop and Shared Task 2019. doi: 10.18653/v1/W19-5034.

[4] Jimeno-Yepes AJ, McInnes BT, Aronson AR. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. BMC Bioinformatics. 2011 Jun 2;12:223. doi: 10.1186/1471-2105-12-223.

[5] Köpcke F, Prokosch HU. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res. 2014 Jul 1;16(7):e161.

[6] Köpcke F, Lubgan D, Fietkau R, Scholler A, Nau C, Stürzl M, Croner R, Prokosch HU, Toddenroth D. Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. BMC Med Inform Decis Mak. 2013 Dec 9;13:134.

[7] Mate S, Bürkle T, Kapsner LA, Toddenroth D, Kampf MO, Sedlmayr M, Castellanos I, Prokosch HU, Kraus S. A method for the graphical modeling of relative temporal constraints. J Biomed Inform. 2019 Dec;100:103314.

[8] McInnes B. An Unsupervised Vector Approach to Biomedical Term Disambiguation: Integrating UMLS and Medline. Proceedings of the ACL-08: HLT Student Research Workshop, Columbus, Ohio. 2008.

[9] Festag S, Spreckelsen C. Word Sense Disambiguation of Medical Terms via Recurrent Convolutional Neural Networks. Stud Health Technol Inform. 2017;236:8-15.