Evaluation of Word Embedding Vector Averaging Functions for Biomedical Word Sense Disambiguation

Rui Antunes and Sérgio Matos

DETI/IEETA, University of Aveiro, 3810-193 Aveiro, Portugal
{ruiantunes,aleixomatos}@ua.pt

Abstract. The biomedical lexicon contains a large amount of term ambiguity, which hinders correct identification of concepts and reduces the accuracy of semantic indexing and information retrieval tools.
Previous work on biomedical word sense disambiguation has shown that supervised machine learning leads to better results than knowledge-based approaches. However, machine learning approaches require the availability of sufficient training data, and generalization performance behind the test data is not known. Knowledge-based methods on the other hand make use of existing knowledge-bases and are therefore mostly limited to the quality of such sources of information about concepts.
In this work, we used word embedding vectors to complement the knowledge-base information. We represent the context of an ambiguous term by the average of the embedding vectors of words around the term, and evaluate the impact of using word distance for weighting this average. We show how this weighting improves the disambiguation accuracy of the knowledge-based approach in a subset of the reference MSH WSD data set from 86% to 88%.

Keywords: biomedical word sense disambiguation, knowledge-based approaches, word embeddings

1 Introduction

Nowadays there is a huge amount of textual data. In order to keep up with all that knowledge, automatic text mining systems for extracting and retrieving information are mandatory. Several Natural Language Processing (NLP) steps have to be accomplished for properly extracting information from text. The most important step is Named Entity Recognition (NER) [1], which deals with the identification of concepts and associates them to knowledge sources, since the whole information extraction task is strongly dependent in the accuracy of the identified concepts. Due to the ambiguity of the human natural language, textual documents are replete of ambiguities, and so the Word Sense Disambiguation (WSD) is a crucial part of the NER task [2]. Particularly in biomedical texts much more terms have many meanings, making it harder to extract accurate information.

Biomedical concept disambiguation aims to remove ambiguity from biomedical documents. Its goal is to normalize the polysemic terms, attributing only one meaning to each ambiguous concept. Given the possible meanings of each ambiguous concept, the surrounding context is used to help inferring the correct meaning. Supervised machine learning techniques and Knowledge-Based (KB) approaches can be followed to solve the WSD problem [3]. Supervised learning algorithms currently achieve the best results, however they require annotated training data. On the other hand, KB approaches have also drawn wide interest [4], since these approaches are less dependent on training data, being strongly dependent on the quality of the knowledge sources. Moreover, the use of multiple knowledge databases has been proven to bring benefits to the problem of WSD [5].

Mikolov et al. [6] proposed a continuous-bag-of-words model architecture for deriving vector representations of words from large unlabeled corpora. These vector representations of words are known as neural word embeddings, or simply word embeddings. This is a recent technique that has been extensively used in several NLP tasks, namely for the WSD task [7, 8].

In the reference biomedical ambiguity MSH WSD data set [9], Jimeno Yepes [10] proposed a supervised biomedical WSD method using word embeddings, achieving a top accuracy around 96% with a Support Vector Machine (SVM) classifier. On the other hand, Sabbir et al. [11] proposed a KB approach that also uses word embeddings, achieving a best accuracy of 92% on the same data set.

In a previous work [12], we proposed a KB method for WSD achieving a top accuracy around 85% in a subset of the MSH WSD data set. In this work, we improved our previous KB method using word distances to weight word embeddings, achieving a best accuracy around 88% in the same subset. We calculate an embedding vector for representing each surrounding context of the ambiguous terms, making a weighted average of the word embeddings using different averaging functions. For the best of our knowledge, Iacobacci et al. [13] were the first to weight word embeddings according to its word distance relative to the ambiguous term. The word embeddings were calculated from around 15 million MEDLINE abstracts. Furthermore, we calculate embeddings vectors for concept textual definitions extracted from the Unified Medical Language System (UMLS) Metathesaurus [14], which were used to calculate cosine similarities between them and the embedding vectors of the surrounding contexts of the ambiguous terms. Association values between two concepts, calculated using the co-occurrences of MeSH terms in MEDLINE citations, were used to weight the cosine similarities. With this approach we were able to infer the most similar sense given the surrounding context of a specific ambiguous term.

2 Biomedical Ambiguity Data Set

Jimeno Yepes et al. [9] proposed a method to automatically develop a biomedical ambiguity data set using the UMLS Metathesaurus and the MeSH indexing of MEDLINE abstracts. Using this method, they created the MSH WSD data set, which is currently the most used data set for evaluating biomedical WSD systems. The MSH WSD data set is composed by 203 biomedical ambiguous entities. Each possible meaning of the ambiguous terms has at maximum 100 instances, where each instance corresponds to a MEDLINE abstract containing the ambiguous term. From the 203 ambiguous terms, 189 only have two possible senses, 12 have three possible senses, and the remaining 2 terms have four and five possible meanings.

In this work we only considered a subset (191 terms) of the MSH WSD data set (the same as in [12]), since some terms[1] have meanings, here represented as Concept Unique Identifiers (CUIs), that we were not able to extract a textual definition from the UMLS 2012 database. The knowledge-based results are considered without using these terms.

3 Knowledge-based Approach

3.1 Word Embeddings

We calculated the word embeddings from around 15 million MEDLINE abstracts. The continuous-bag-of-words model architecture [6] was used to generate the word embedding models using the Gensim framework [15] implemented in Python. These generated word embeddings were used to calculate embedding vectors for the CUI textual definitions, and for the surrounding context of each ambiguous term.

3.2 Word Embedding Averaging Functions

In the first place, Inverse Document Frequency (IDF) values were calculated using the CUI textual definitions, where each definition represented a document. We used the IDF formula that is expressed in (1), where N is the total number of documents, and df_t is the document frequency of the term t.

$$\mathrm{IDF}(t) = \log_{10} \frac{N}{df_t} \tag{1}$$

Afterwards, the embedding vectors of the CUI textual definitions were calculated using the Term Frequency – Inverse Document Frequency (TF-IDF) weighting scheme. And on the other hand, the embedding vectors of the surrounding contexts of the ambiguous terms were weighted using the pre-calculated IDF values and a word distance function $\mathrm{f}(d)$.

Distinct word embedding averaging functions were defined for being applied only in the abstracts containing the ambiguous terms. For each word w in the surrounding context of an ambiguous term t, we used its absolute word distance $\mathrm{d}(w, t)$ to calculate a weighted context vector. The absolute word distance $\mathrm{d}(w, t)$

[1] Terms not used: Ca; CNS; Crown; DBA; FAS; Gamma-Interferon; Hybridization; ITP; PCP; Plaque; Pneumocystis; Semen.

is used as input parameter for a specific decay function $\mathrm{f}(d)$. The objective was to give a greater importance to words closest to the ambiguous term. The weighted calculus of the context embedding vector is shown in (2), where the function $\mathrm{WE}(w)$ represents the embedding vector of a specific word w. All the calculated embedding vectors for the CUIs and the contexts were normalized.

$$\text{embedding\_vector}(context) = \sum_{w \in context} \mathrm{IDF}(w) \cdot \mathrm{f}(\mathrm{d}(w,t)) \cdot \mathrm{WE}(w) \qquad (2)$$

Four decay functions were defined and tested for weighting the calculus of the context embedding vectors:

- No decay: $\mathrm{f}(d) = 1$;
- Fractional decay: $\mathrm{f}(d) = 1/d$;
- Exponential decay: $\mathrm{f}(d) = \exp(-d)$;
- Logarithmic decay: $\mathrm{f}(d) = 1/\ln(1+d)$.

3.3 Method

Firstly, CUI textual definitions were extracted from the UMLS database, and each CUI was represented as an embedding vector calculated as a weighted average of the word embeddings with the TF-IDF scheme. Each abstract of the MSH WSD data set, that is, each surrounding context of the ambiguous terms were also mapped to an embedding vector using different word embedding averaging functions as described before.

For each abstract, we could calculate the Cosine Similarity (CS) between the embedding vector of the surrounding context and the embedding vector of each possible meaning (specified by a CUI) to select the CUI with highest cosine similarity as the correct meaning. However, we extended this baseline approach calculating cosine similarities between the context embedding vector and each of the CUI definition vectors. And for weighting these cosine similarities we used CUI-CUI association values that were calculated as normalized Pointwise Mutual Information (nPMI) values from the co-ocurrence of MeSH terms in MEDLINE citations[2]. The score for each possible CUI of an ambiguous term is then calculated as shown in (3), where the CUI that achieves the highest score is inferred as the correct sense.

$$\mathrm{score}(CUI) = \frac{1}{N} \sum_{j} \mathrm{nPMI}(CUI, CUI_j) \cdot \mathrm{CS}(\boldsymbol{t}, \boldsymbol{CUI_j}) \qquad (3)$$

According to (3), for each possible CUI the cosine similarities between its context embedding vector $\boldsymbol{t}$ and the concept embedding vectors $\boldsymbol{CUI_j}$ are weighted by the respective nPMI values. The nPMI value of a CUI in relation to himself has a value of a unit. However, some nPMI association values are undefined,

[2] https://ii.nlm.nih.gov/MRCOC.shtml.

since the nPMI values were restricted using different thresholds. The final division of the score by N is a normalization, since N is the total number of nPMI associations used to weight the N cosine similarities.

4 Results

Simulations with different nPMI thresholds $(0.2, 0.3, \ldots, 1.0)$ were performed. Table 1a shows the results for nPMI $= 1$, that is the case when only the cosine similarity between the context vector and the CUI vector is considered. On the other hand, table 1b shows the results for nPMI ≥ 0.3, which was the nPMI threshold that produced the best results. The first row of the table 1b shows the results with no decay function, which produced an accuracy of 86% with the word embedding model of size 100 and window of 50 words. From table 1b one can see that the exponential decay weighting obtained the lowest results even when compared to using no decay. The fractional decay produced the best results, achieving an accuracy of 85% with nPMI $= 1$, and a best accuracy of 88% with nPMI ≥ 0.3. Overall, the use of the concept association values nPMI ≥ 0.3 allowed to improve the overall accuracies around 3% when compared to not using any related concepts, that is, the case when only the cosine similarity between the CUI definition and the context of the ambiguous concept was considered.

Table 1. Knowledge-based accuracies using four different word embedding averaging functions with IDF weighting. f(d): word embedding averaging function, where d is the absolute distance to the ambiguous term; S: size; W: window; nPMI: normalized Pointwise Mutual Information.

(a) nPMI $= 1$: only the cosine similarity between the context embedding vector and the concept definition vector is used.

f(d)	S100			S300		
	W5	W20	W50	W5	W20	W50
1	0.8078	0.8194	0.8200	0.8077	0.8194	0.8182
$1/d$	0.8375	0.8468	0.8461	0.8408	**0.8471**	0.8456
$\exp(-d)$	0.8227	0.8255	0.8233	0.8262	0.8257	0.8214
$1/\ln(1+d)$	0.8216	0.8347	0.8361	0.8232	0.8344	0.8343

(b) nPMI ≥ 0.3: related concepts with a nPMI value higher than 0.3 are the ones considered to weight the cosine similarities.

f(d)	S100			S300		
	W5	W20	W50	W5	W20	W50
1	0.8430	0.8585	0.8604	0.8428	0.8552	0.8541
$1/d$	0.8638	0.8766	**0.8795**	0.8641	0.8751	0.8747
$\exp(-d)$	0.8418	0.8522	0.8515	0.8421	0.8505	0.8487
$1/\ln(1+d)$	0.8539	0.8686	0.8717	0.8539	0.8674	0.8678

5 Conclusions

We showed that adding a word distance weighting in the calculus of the context embedding vector improved the accuracy of our proposed KB method in 2%, achieving a best accuracy around 88% using the fractional decay. This result is slightly above the results obtained with the three KB methods proposed by [9] which achieved a top accuracy around 84%. Tulkens et al. [16] also applied a similar KB method to the same data set using BioASQ word embeddings, obtaining a disambiguation accuracy of 84%. They also compared UMLS definitions with the contexts of the ambiguous terms. Sabbir et al. [11] used a KB approach with neural concept embeddings and distant supervision, achieving a top state-of-the-art accuracy around 92%.

One restraint of our method is that because we were not able to extract definitions for some CUIs, the disambiguation could not be performed to all terms of the data set. As future work, we intend to overcome this problem searching for textual definitions in other databases. Also, we intend to apply this word embedding distance weighting in a supervised learning approach to verify the veracity of this method.

Acknowledgements. This work was supported by Portuguese national funds through FCT - Foundation for Science and Technology, in the context of the project IF/01694/2013/CP1162/CT0018. Sérgio Matos is funded under the FCT Investigator Programme.

References

1. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticæ Investigationes **30**(1) (2007) 3–26
2. Navigli, R.: Word sense disambiguation: a survey. ACM Computing Surveys **41**(2) (2009) 10:1–10:69
3. McInnes, B.T., Stevenson, M.: Determining the difficulty of word sense disambiguation. Journal of Biomedical Informatics **47** (2014) 83–90
4. Garla, V.N., Brandt, C.: Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. Journal of the American Medical Informatics Association **20**(5) (2013) 882
5. Tsai, C.T., Roth, D.: Concept grounding to multiple knowledge bases via indirect supervision. Transactions of the Association for Computational Linguistics **4** (2016) 141–154
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv e-print (2013)
7. Wu, Y., Xu, J., Zhang, Y., Xu, H.: Clinical abbreviation disambiguation using neural word embeddings. In: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), Beijing, China, Association for Computational Linguistics (2015) 171–176
8. Taghipour, K., Ng, H.T.: Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In: Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT 2015), Denver, Colorado, USA (2015) 314–323

9. Jimeno-Yepes, A., McInnes, B.T., Aronson, A.R.: Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. BMC Bioinformatics **12**(1) (2011) 223
10. Jimeno-Yepes, A.: Word embeddings and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation. arXiv e-print (2016)
11. Sabbir, A.K.M., Jimeno-Yepes, A., Ramakanth, K.: Knowledge-based biomedical word sense disambiguation with neural concept embeddings and distant supervision. arXiv e-print (2017)
12. Antunes, R., Matos, S.: Biomedical word sense disambiguation with word embeddings. In: 11th International Conference on Practical Applications of Computational Biology & Bioinformatics. Springer International Publishing (2017) 273–279
13. Iacobacci, I., Pilehvar, M.T., Navigli, R.: Embeddings for word sense disambiguation: an evaluation study. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, Association for Computational Linguistics (2016) 897–907
14. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research **32**(Suppl 1) (2004) D267
15. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on NewChallenges for NLP Frameworks, Valletta, Malta (2010) 45–50
16. Tulkens, S., Šuster, S., Daelemans, W.: Using distributed representations to disambiguate biomedical and clinical concepts. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, Berlin, Germany, Association for Computational Linguistics (2016) 77–82