A One-size-fits-all Bidirectional LSTM for Word Sense Disambiguation of Biomedical Text Data

1 Motivation

Natural Language Processing (NLP) techniques, e.g., named entity recognition algorithms [1], syntactic parsers [2], and relation extractors [3], build the foundation of many high-level biomedical information extraction and knowledge discovery systems [4, 5]. If any fault in the foundation, there would be a domino effect of errors in initial components leading to more faults in the higher ones. Biomedical text data often abound with ambiguous terms that introduce further challenges. That is, the full system, including its components in various levels, will suffer from the ambiguity problem if it is not resolved properly.

2 Problem Statement

Word Sense Disambiguation (WSD) attempts to predict the correct sense of an ambiguous term within a *context*, given a set of candidates. For example, consider the ambiguous term *Ca* in the following sentence "*Ca intakes in the United States and Canada appear satisfactory among young adults.*" The term *Ca* holds the following candidate senses, *Canada* (s_1), *California* (s_2), *Calcium* (s_3), and *Cornu ammonis* (s_4). An accurate WSD algorithm must predict the correct sense s_3 for that sentence.

From the NLP perspective, two main challenges exist when we address WSD, which are the *lexical sample* task and the *all-words* task. While supervised algorithms have shown promising results for the lexical sample task, unsupervised methods outperformed them for the more challenging task of all-words. This is due to the fact that the supervised algorithms typically build *one separate classifier* for *each* ambiguous term, which is *exclusively* trained on instances of that term. That is, for an ambiguous term, a *softmax* layer with a *cross entropy* or *hinge loss* is often determined to parametrize the corresponding weight matrix and bias vector with respect to senses of the term. Hence, they suit the lexical sample task well, but not the all-words task for which a WSD algorithm should work well *collectively*, i.e., it must take all ambiguous terms into account, not just a few selected set of them. That means, for the current supervised algorithms, a large number of annotated instances are needed to train an accurate WSD model for each ambiguous term [6, 7]. To alleviate these challenges, we propose the deepBioWSD network in the next section.

3 Approach: One-size-fits-all Bidirectional LSTM

We introduce deepBioWSD, as a one-size-fits-all network that uses a Bidirectional Long Short-Term Memory network (BLSTM) in its architecture, and works with neural *sense embeddings*. The sense embeddings may be pre-trained, in either a supervised or an unsupervised way, in advance. Supervisedly, our network is trained on *all* instances of ambiguous terms as one group, while *sense-context pairs* and $s_i \in \{0.0, 1.0\}$ constitute the input and the output of the network, respectively. For that, the deepBioWSD network shares parameters, i.e., weights and biases, over all senses, of all terms, in contrast to the existing (supervised) algorithms, in which different parameters are exclusively learned for the ambiguous terms. While being computationally efficient, such structure helps to encode statistical information across all (ambiguous) terms and enables the network to predict the correct sense of every term in a context - or even to offer semantically/syntactically alternative answers for a blank space in a text. Given a document and the position of a target term, the deepBioWSD network computes a probability distribution over possible senses associated with that term.

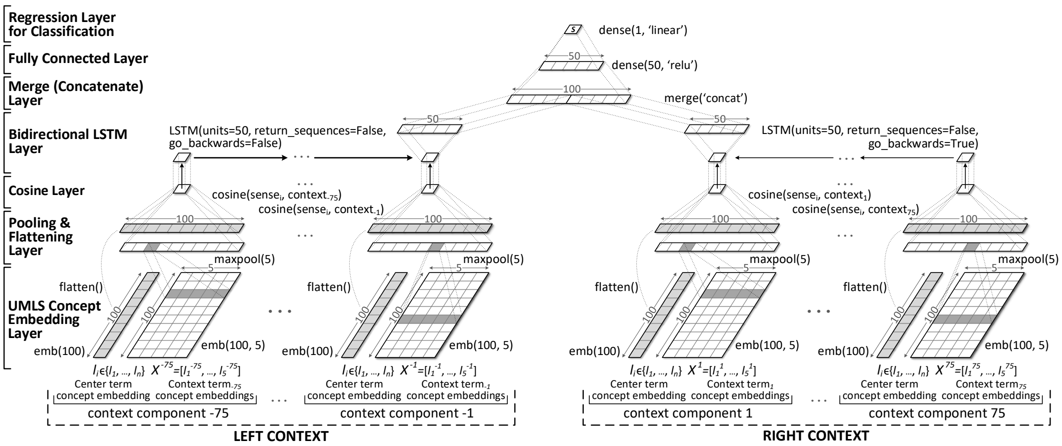
Regression Layer for Classification
Fully Connected Layer
Merge (Concatenate) Layer
Bidirectional LSTM Layer
Cosine Layer
Pooling & Flattening Layer
UMLS Concept Embedding Layer
s
dense(1, 'linear')
50
dense(50, 'relu')
100
merge('concat')
50
LSTM(units=50, return_sequences=False, go_backwards=False)
50
LSTM(units=50, return_sequences=False, go_backwards=True)
cosine(sense$_i$, context$_{-75}$)
cosine(sense$_i$, context$_{-1}$)
cosine(sense$_i$, context$_1$)
cosine(sense$_i$, context$_{75}$)
100
maxpool(5)
5
flatten()
100
emb(100)
emb(100, 5)
$l_i \in \{l_1, ..., l_n\}$
$X^{-75} = [l_1^{-75}, ..., l_5^{-75}]$
$X^{-1} = [l_1^{-1}, ..., l_5^{-1}]$
$X^{1} = [l_1^{1}, ..., l_5^{1}]$
$X^{75} = [l_1^{75}, ..., l_5^{75}]$
Center term concept embedding
Context term$_{-75}$ concept embeddings
Context term$_{-1}$ concept embeddings
Context term$_1$ concept embeddings
Context term$_{75}$ concept embeddings
context component -75
context component -1
context component 1
context component 75
LEFT CONTEXT
RIGHT CONTEXT

Figure 1: The deepBioWSD Network

As illustrated in Fig. 1, our network consists of seven layers. Due to the replacement of the conventional softmax layer with a *linear* (regression) layer on the top of our network, we imposed a modification in the input layer. That is, not only the contextual features are going to make an input of the network, but also, the sense for which we are interested to find whether the given context is meaningful is also provided as an input. For an ambiguous term with the sense-set $\{s_1, \ldots, s_n\}$, after computing cosine similarities of every sense with the senses of the context terms, we expect the sequence result of the cosine similarities between the correct sense and the surrounding context communicate a pattern-like information that can be encoded by a BLSTM network; however, for the incorrect senses this premise does not hold. The reader may note that several studies already incorporated the idea of sense-context cosine similarities in their WSD models [8, 9]. Nonetheless, the context terms, which are determined by the SPECIALIST Lexicon during the disambiguation process, can be ambiguous themselves. To deal with their ambiguity, just before the cosine layer, a pooling layer is devised, the result of which attempts to estimate the sense of the ambiguous terms appeared in the context. Our design makes the network to take gradients with respect to (shared) sense embeddings of both the target term and the context terms simultaneously.

Assuming $\hat{y}_{s_i}$ as the output, during network training, for an instance with its given context and the correct sense as inputs, $\hat{y}_{s_i}$ is set to be **1.0**, whereas for the same context with incorrect senses it is set to be **0.0**. During testing however, among all the senses, the output of the network for a sense that gives the highest value of $\hat{y}_{s_i}$ is considered as the true sense of the ambiguous term. In other words, the predicted sense is:

$$\underset{s_i}{\operatorname{argmax}}\{\hat{y}_{s_1}, ..., \hat{y}_{s_n}\}, \quad s_i \in \{s_1, ..., s_n\} \tag{1}$$

By applying softmax to the results of the estimated values $\{\hat{y}_{s_1}, ..., \hat{y}_{s_n}\}$, we can represent them as probabilities. This will facilitate interpretation of them especially when deepBioWSD is benefiting from an *active learning* setting where intricacy and importance of one instance can be measured.

4 Discussion and Conclusion

Evaluating on the MSH-WSD dataset [10], we observed that deepBioWSD outperformed the state-of-the-art methods. Our one-size-fits-all network, though demanding less number of training data, achieved an accuracy which was notably higher than the accuracy of the recent deep learning-based one-network-per-one-term approaches. That merits deepBioWSD to be conveniently deployable in real-time environments.

To conclude our observations, the deepBioWSD network particularly found two types of instances challenging. First, when the syntactic structure surrounding candidate senses were very similar (e.g., *Switzerland* and *China* for the ambiguous term *CH*). Second, when the senses are (semantically) too close so much that they share the same immediate parent in the UMLS, or one term directly subsumes the other sense (as in *lymphogranulomatosis*).

References

[1] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

[2] Sahil Garg, Aram Galstyan, Ulf Hermjakob, and Daniel Marcu. Extracting biomolecular interactions using semantic parsing of biomedical text. In *AAAI*, pages 2718–2726, 2016.

[3] Kyubum Lee, Sunwon Lee, Sungjoon Park, Sunkyu Kim, Suhkyung Kim, Kwanghun Choi, Aik Choon Tan, and Jaewoo Kang. Bronco: Biomedical entity relation oncology corpus for extracting gene-variant-disease-drug relations. *Database*, 2016, 2016.

[4] Ahmad P Tafti, Jonathan Badger, Eric LaRose, Ehsan Shirzadi, Andrea Mahnke, John Mayer, Zhan Ye, David Page, and Peggy Peissig. Adverse drug event discovery using biomedical literature: a big data neural network adventure. *JMIR medical informatics*, 5(4), 2017.

[5] Kyubum Lee, Wonho Shin, Byounggun Kim, Sunwon Lee, Yonghwa Choi, Sunkyu Kim, Minji Jeon, Aik Choon Tan, and Jaewoo Kang. Hipub: translating pubmed and pmc texts to networks for knowledge discovery. *Bioinformatics*, 32(18):2886–2888, 2016.

[6] Yue Wang, Kai Zheng, Hua Xu, and Qiaozhu Mei. Clinical word sense disambiguation with interactive search and classification. In *AMIA Annual Symposium Proceedings*, volume 2016, page 2062. American Medical Informatics Association, 2016.

[7] Mohammad Taher Pilehvar and Roberto Navigli. A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. *Computational Linguistics*, 40(4):837–881, 2014.

[8] Bridget T. McInnes and Ted Pedersen. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of biomedical informatics*, 46(6):1116–1124, 2013.

[9] Ahmad Pesaranghader, Ali Pesaranghader, Stan Matwin, and Marina Sokolova. One single deep bidirectional lstm network for word sense disambiguation of text data. In *Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8–11, 2018, Proceedings 31*, pages 96–107. Springer, 2018.

[10] Antonio Jimeno-Yepes and Bridget T. McInnes. MSH WSD dataset. `https://wsd.nlm.nih.gov/collaboration.shtml`, 2015. (Last access 12-July-2018).