Jestr

Journal of Engineering Science and Technology Review 12 (6) (2019) 105 - 112

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Disambiguation of Biomedical Acronyms Based on a Bidirectional Recurrent Neural Network of Character-level Features

Ren Kai[1,*], Li Na[1], Xiong Wei[1] and Wang Shi-Wen[2]

[1]*College of Computer Science, South-Central University for Nationalities, Wuhan, China*
[2]*Sociologue, Université Toulouse-Jean-Jaurès, Toulouse, France*

Received 28 May 2019; Accepted 13 December 2019

Abstract

Polysemic acronyms are very common in the field of biomedicine. These acronyms have different senses in different contexts. The ambiguity of acronyms may cause significant negative impact on the understanding of the full text by machine learning. To address the disambiguation of acronyms in the biomedical domain, most associated studies are based on methods using word-level contextual features. These methods require abundant relevant external resources for model training, and the accuracy of their disambiguation of acronyms may decrease greatly upon the lack of external resources. In this study, disambiguation of biomedical acronyms was investigated on the basis of the character-level feature model to realize the disambiguation of biomedical acronyms with largely limited external corpora. First, sentences containing ambiguous acronyms were extracted through retrieval and the feature vector of the context were initialized by using the character-level features. Second, these initial vectors were input into the bidirectional long short-term memory neutral network model for training. Finally, the disambiguation of acronyms was realized by the outputs of the neutral network model through the Softmax classification approach. The results of acronym disambiguation based on character-level feature model were also compared with those based on word-level feature models. Results demonstrate that the average accuracy of the character-level feature neutral network algorithm reaches 85.82% on the dataset of 106 common biomedical acronyms. Thus, the character-level feature neutral network algorithm is superior to the traditional methods, which use a large number of external resources. This study confirms that the disambiguation method based on character-level features is applicable to the disambiguation of biomedical acronyms under limited relevant data.

Keywords: WSD, Bi-LSTM, Biomedical, Abbreviation

1. Introduction

Bioinformatics explores effective information from mass biomedical information by using computer and informatics technology, thereby providing assistance to clinical medicine. In the biomedical field, "Synonym" and "Homograph" phenomena are common. These problems are known as disambiguation problems in the natural language processing (NLP) field. Despite the polysemy of ordinary medical terms, a special type of polysemy phenomena is described, that is, the polysemy of acronyms. In biomedical studies, many important concepts and terms are expressed in acronyms. For example, the acronym AA can mean either amino acids or alcoholics anonymous. The acronym RA can be refractory anemias, radium, or rheumatoid arthritis. Hence, the accurate disambiguation of acronyms using a computer is important to understand and analyze biomedical data [1].

Acronyms are used throughout in biomedical studies, except for the full term in the first occurrence. Some biomedical acronyms have an explicit sense in a specific context, and they are often used directly in many biomedical studies instead of the full terms. The specific sense of acronyms has to be judged by experience according to context. Given that the automatic computer processing of texts lacks the common medical knowledge of doctors [2], choosing the accurate sense of ambiguous acronyms is difficult. Some scholars [3] pointed out that conventional disambiguation methods of biomedical acronyms have low accuracy and often require extensive external annotated corpus [4].

To address the high dependence of disambiguation of biomedical acronyms on large-scale external resources, a disambiguation method based on character-level feature model was designed in this study. This method requires no abundant external resources. A comparative study between the proposed method and the current disambiguation methods based on the embedding of words learned from large-scale text was carried out. This study realized the disambiguation of biomedical acronyms with limited external resources.

2. State of the art

Disambiguation of words often requires a certain amount of manually annotated corpus to train the model. However, manually annotated corpus can only be acquired by consuming considerable manpower and time, which are impossible in an extensive mode. To address these problems, some scholars have attempted to train the model by introducing it in external public resources and graph theory. These methods require no extra manually annotated data or only need few manually annotated data to realize

*E-mail address: rk8123@gmail.com
ISSN: 1791-2377
doi:10.25103/jestr.126.13

disambiguation. Nevertheless, such methods are often inferior to those using manually annotated data. Recently, some representative disambiguation methods based on external resources have been developed. Henry [5] implemented the disambiguation of biomedical words based on knowledge by using a feature extraction method, which achieved good effect. Dongsuk [6] expressed vectors of words based on knowledge map and thereby gained correlation between different words, thereby accomplishing disambiguation. Duque [7] accomplished the disambiguation of biomedical words based on the co-occurrence graph of context. For disambiguation of biomedical words, the standard linguistic dataset MeSH Word Sense Disambiguation (MSH-WSD) constructed by Jimeno [8] for the disambiguation of biomedical words was applied. The MSH-WSD dataset offers a unified evaluation standard for subsequent studies. In the subsequent studies [9] on disambiguation based on external knowledge sources, disambiguation of words was implemented by combining Automatic Extracted Corpus (AEC) and Machine Readable Dictionary (MRD) with collocation features. Moreover, the disambiguation effect was evaluated on the MSH-WSD. All these methods have achieved good effects, but they have a common problem of heavy dependence on external knowledge sources. The quantity and quality of external resources can directly influence the disambiguation of words.

With respect to universal excessive dependence on external resources, Pasini [10] proposed a multilingual disambiguation system that does not use manual annotated training data. Panchenko [11] also proposed an unsupervised disambiguation method without the use of external knowledge. Charbonnier [12] and Li [13] carried out in-depth studies on the unsupervised disambiguation of acronyms and proposed a special disambiguation method in accordance with the features of acronyms. The author of the present study studied unsupervised disambiguation of medical terms based on kernel fuzzy C-means clustering [14]. These methods are all disambiguation methods developed from traditional fields, and they have a common problem. Specifically, none of these methods have optimized features of ambiguous acronyms, especially in the biomedical domain, thus resulting in poor performance.

Previous studies on text expression focused on word-level features. Recently, some scholars discovered that character-level features made special contributions to the expression of textual features. Some scholars [15] [16] tried to add character features into word features and achieved breakthroughs in classification and machine translation. These studies proved the importance of character features in text expression. However, no studies on the disambiguation of biomedical acronyms based on character-level features have been reported yet.

With the increasing application of deeplearning technology in the disambiguation of word sense, Henry [17] carried out an in-depth study on the feature selection of biomedical contexts. Wang [18] extended the word features into sentence-level features and combined them in a further study. Liu [19] and Kumar [20] focused on the sense of words and studied word embedding. Such reverse thinking provided new insights into disambiguation. Ragangato [21] attempted to apply the neutral network model into disambiguation in the traditional field. Le [22] investigated the performances of the long short-term memory (LSTM) neutral network model in the disambiguation of word sense. These methods based on deep learning technology in the traditional fields have a principal problem—using large data in the relevant text for word embedding training in advance. The whole disambiguation process also requires support from abundant external data and computing resources. Given these features, such algorithms perform poorly in biomedical fields with resource shortages or high requirements on computing timeliness.

Overall, previous studies on the disambiguation of acronyms in the biomedical field generally used the method of word-level features. Although these methods have good relative effects, they often require abundant external resources to train the model. In this study, disambiguation of acronyms was accomplished using deep neutral network based on character-level features. This method does not require word training in advance, and it can adapt to various environments with incomplete external data. Meanwhile, the proposed disambiguation method was compared with deep neutral network method based on word-level features. This study is expected to determine the best disambiguation method of biomedical acronyms independent of external resources.

The remainder of this study is organized as follows: Section 3 introduces the model and method applied in this study, Section 4 presents the experimental design and analysis of results, and Section 5 states the conclusions.

3. Methodology

3.1 Character-level model

The character-level model is different from other models. It uses character as the minimum unit, whereas other traditional models use word as the minimum unit. Generally, choosing an English word as the minimum unit is reasonable. When an English word is further divided into independent letters, the correlation of different letters is not so close with that of different words. According to the specific disambiguation tasks of biomedical acronyms, acronyms are generally composed of several characters, and the full terms of these acronyms are extended from each letter. When the acronym is complete, it may have different full terms. The correlation among characters in an acronym is believed to be stronger than that among characters in traditional words. Therefore, this study attempted to realize the disambiguation of biomedical acronyms based on the character-level model.

In the typical disambiguation of biomedical acronyms, several contexts that contain the studied acronyms will be offered first. For example, an ambiguous acronym X is used in the abstract of biomedical studies, and this acronym X has two standard senses: M1 and M2. The disambiguation task is to determine the correct sense of X according to context.

A disambiguation model is constructed. In this model, X(i) is a continuous text that contains the ambiguous word. Under general conditions, it is a word sequence. Y(i) is used to express the senses of the ambiguous acronym in X(i). Briefly, Y(i) is M1 or M2. The model constructs a network $f = X(i)$ to predict the sense tag of the sample.

In traditional models, the text is generally divided into a series of words. Bag-of-word, n-gram, and word-based neutral network models are used for these words. Later, texts are classified according to the prediction results of the model by learning these word-based feature vectors. On this basis, the correct sense of the ambiguous acronym in context can be determined according to classification results. This approach is the typical disambiguation model based on word-level features.

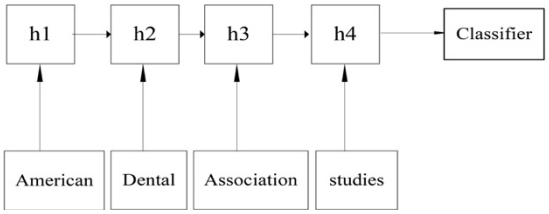

Fig. 1. Classification model based on word-level features

The recurrent neural network (RNN) classification model based on word-level features is shown in Fig. 1. Each word is encoded as a vector, and the produced vectors contribute classification results through different classification models in the next stage. The classification results can be used to disambiguate directly.

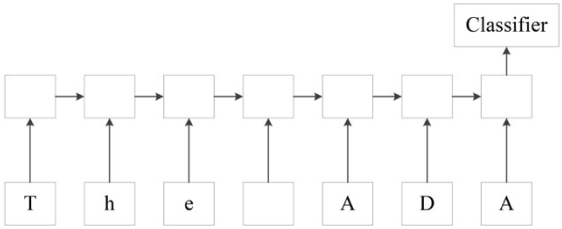

Fig. 2. Classification model based on character-level features

Character-level model is a model with fine-grained features compared with the word model. Fig. 2 shows that a classification model of character-level RNN reads characters in order one by one, and a window with a fixed length can be selected. The embedding vector of context is constructed by using characters in the window, and the generated character-based vectors are used as the input of the neutral network.

Neutral network model based on character-level features has certain advantages to retain context features in principle. In acronyms, each character is the minimum unit with a specific meaning. If acronyms can form a word as an integral, the semantic features of each independent letters will be ignored. This character-level feature model is the model applied in the experiment in Section 4.

3.2 One-Hot encoding

One-Hot encoding is a binary coding vector that is often expressed by a 1D vector matrix in NLP.

In One-Hot encoding, which uses character as the minimum unit, all characters include the capital and small letters of the 26 letters and the symbols of finite categories, such as punctuation marks. Generally, preprocessing is performed before encoding to eliminate unnecessary characters and unify the capital or small letters. After processing, the length of One-Hot encoding is fixed, and all characters are combined to form the input vectors.

In One-Hot encoding, the characters in a context are used as the input, and each character is input into the subsequent model as One-Hot form. In common English text, approximately 70 characters are found, including the capital and small letters of 26 letters, 10 digital characters, and 33 other characters, such as punctuation marks and common symbols on the keyboard. In the One-Hot model, the vector length can be determined and can be decreased by preprocessing.

The generation character feature vectors are based on the One-Hot model of the characters. The major “end-to-end” model is shown in Fig. 3. In a character feature model, other artificial features, such as part-of-speech, Brown, and parsing features, are neglected. The model uses character set as the input simply to accomplish feature extraction, model learning, and conclusion generation, thereby truly eliminating artificial interventions.

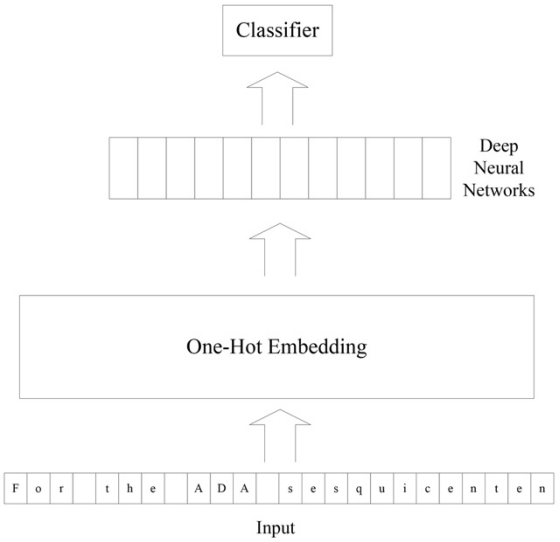

Fig. 3. End-to-end disambiguation of word sense based on the One-Hot model of characters

The One-Hot model may cause a curse of dimensionality when the input data size is too large, thereby decreasing the training efficiency. In principle, complete feature information of the model based on characters is available, which will surely increase the complexity of computation. Instead, an equilibrium selection can only be made according to practical uses.

In Section 4, the One-Hot encoding that uses character as the basic units was applied as the input data. Data were placed in a neutral network model to realize the disambiguation of biomedical acronyms.

3.3 Bidirectional recurrent neutral network

In context feature leaning, information surrounding the keywords is often the most important. The RNN can only visit the past context information. To solve this problem, it has to extend visits to future context to learn relative context features.

Given that standard bidirectional recurrent neutral network (BRNN) cannot master future information on a time sequence, the new RNN adds a delay memory process between the input and the goal, thereby increasing both past and future context features. The improved algorithm generally adds future context information at multiple time points and the past context to predict the output.

Fig. 4 shows that the forward and backward of each training sequence in a BRNN are two RNNs, and each RNN connects an output layer. A non-connecting relationship is set between the forward and backward hidden layers in the model. Among six weights in the extended BRNN model based on time sequence in Fig. 4, w1 and w2 are the parameters from the input to the forward and backward hidden layers, w2 and w5 are the parameters from the hidden layer to the hidden layer, and w4 and w6 are the parameters from the forward and backward hidden layers to the output layer. The LSTM model is gained by improving this model.

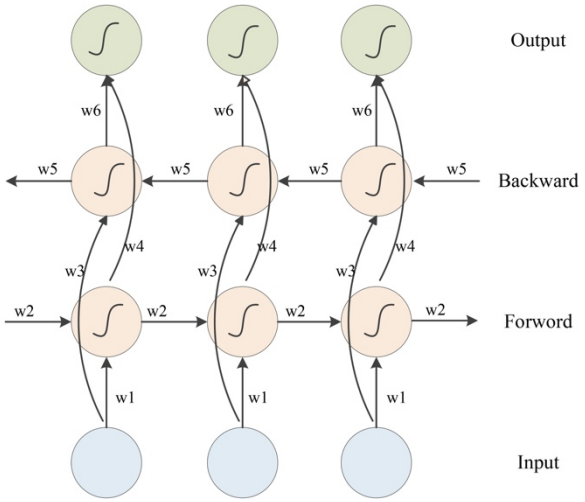

Fig. 4. Temporal extension of BRNN

3.4 Bidirectional long short-term memory

The mapping process of BRNN between the input and output sequences contains context information. However, a standard BRNN has a limited range of context information storage. The effects of a hidden layer on network output may decline gradually. A bidirectional long short-term memory (Bi-LSTM) structure is applied to solve this problem by embedding an LSTM module into an RNN. The details of this LSTM module are shown in Fig. 5.

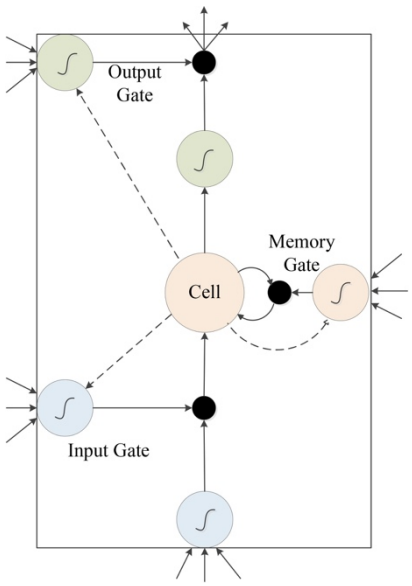

Fig. 5. LSTM module

The LSTM module is divided into forward layer and backward layer. The related process is introduced as follows. The input gate is controlled by the parameters in Eqs. (1) and (2). In Eq. (1), the input network with a length of t have l input units, H hidden units, and C output units. The parameters a in Eq. (1) and b in Eq. (2) reflect the input of the network unit at t and the output of the unit nonlinear differentiable activation function at t, respectively. In Eq. (1), ω is the weight matrix, x is the external output, b is the input of the hidden unit, and s is output of the cell. Eqs. (3)–(9) are similar to Eqs. (1) and (2), which reflect the composition of memory gate, cell, output gate, and output unit module. A complete-sequence hidden unit can be gained from recursive recall.

Input gate:

$$a_l^t = \sum_{i=1}^{I} \omega_{il} x_i^t + \sum_{h=1}^{H} \omega_{hl} b_h^{t-1} + \sum_{c=1}^{C} \omega_{cl} s_c^{t-1} \tag{1}$$

$$b_l^t = f\left(a_l^t\right) \tag{2}$$

Memory gate:

$$a_\phi^t = \sum_{i=1}^{I} \omega_{i\phi} x_i^t + \sum_{h=1}^{H} \omega_{h\phi} b_h^{t-1} + \sum_{c=1}^{C} \omega_{c\phi} s_c^{t-1} \tag{3}$$

$$b_\phi^t = f\left(a_\phi^t\right) \tag{4}$$

Cell:

$$a_c^t = \sum_{i=1}^{I} \omega_{ic} x_i^t + \sum_{h=1}^{H} \omega_{hc} b_h^{t-1} \tag{5}$$

$$s_c^t = b_\phi^t s_c^{t-1} + b_l^t g\left(a_c^t\right) \tag{6}$$

Output gate:

$$a_\omega^t = \sum_{i=1}^{I} \omega_{i\omega} x_i^t + \sum_{h=1}^{H} \omega_{h\omega} b_h^{t-1} + \sum_{c=1}^{C} \omega_{cw} s_c^{t-1} \tag{7}$$

$$b_\omega^t = f\left(a_\omega^t\right) \tag{8}$$

Output unit:

$$b_c^t = b_\omega^t h\left(s_c^t\right) \tag{9}$$

Fig. 5 shows that inputs of the input gate are the external input at t, output of the hidden unit at t-1, and output of the cell at t-1. The inputs of the memory gate are external input at t, output of hidden unit at t-1, and output of cell at t-1. The inputs of unit include the sum of the product between the output of memory gate at t and the output of unit at t-1 as well as the sum of the product between the output of the input gate at t and the activation function. The inputs of the output gate include the external input at t, output of the hidden unit at t-1, and output of the cell at t. The output of the unit is the output of the output gate at t multiplied by the output of cell at t.

Backward layer is very similar to forward layer. This network is called Bi-LSTM RNN. It is equipped with advantages of bidirectional neural network and LSTM. In Section 4, this LSTM neutral network model is applied to the disambiguation of acronyms based on character input or word input.

4. Result Analysis and Discussion

4.1 Experimental environment

In this experiment, the corpus of the National Library of Medicine's MSH-WSD was applied. It has 203 ambiguous words, including 106 acronyms, 88 terms, and 9 hybrid ambiguous words. In this experiment, 106 acronyms were used as corpus.

The character-level test was carried out by using the One-Hot encoding presented in Section 3.2, and the length of the window was set to 200. The character-level model uses no word embedding form. The word embedding model in the control word-level model generated 200D word vector for training the Bi-LSTM model. In this experiment, the crawled data and the MSH-WSD corpus were combined as the training data of the GloVe model for word embedding. The results were used as input for the follow-up Bi-LSTM model.

4.2 Experiment model

The experiment was based on the MSH-WSD corpus. This experiment involved the encoding of text based on character level and finally realizing the disambiguation of acronyms. First, the text was divided into sentences, and only sentences containing ambiguous acronyms were chosen for preprocessing. This step restricts the maximum length of a sequence so the model can accomplish disambiguation under existing computation conditions. Input vectors with fixed length were generated from the One-Hot encoding of context characters and then input into the improved Bi-LSTM neutral network presented in Section 3.4. The final results were output to a classifier, thereby accomplishing the disambiguation of acronyms.

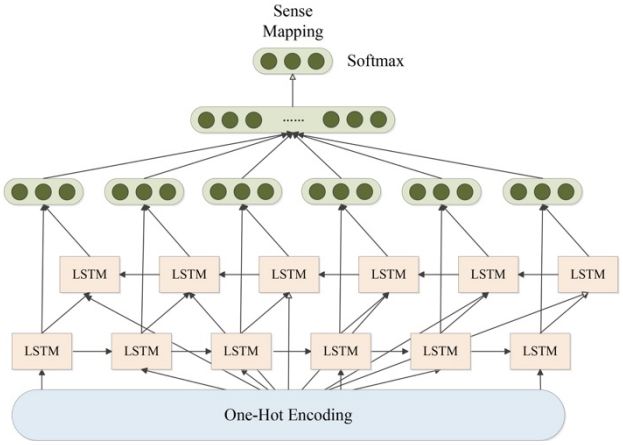

Fig. 6. Disambiguation model based on the Bi-LSTM of characters

The complete model structure is shown in Fig. 6. In this model, character is used as the basic unit, and character order of words in the context is read as a sequence. The model contains front and back characters of the keywords and relevant positional information. It forms context codes based on character-level features by setting window length. The character codes of each context are transmitted to the next Bi-LSTM model layer through the final formation encoding, and the model output is gained after the multiple iterations of the Bi-LSTM layer realizes the model goal. Finally, the output of the Bi-LSTM model is used to classify text by SoftMax function. On the basis of the classification results, each category is mapped onto a similar sense, thereby realizing the disambiguation of acronyms. These steps comprise the whole experimental process.

4.3 Comparative experimental design

The comparative experiment chose AEC and MRD, which were Jimeno-Yepes' disambiguation methods based on external knowledge. These two methods have been applied in the disambiguation of acronyms and evaluated based on the MSH-WSD. In this study, AEC and MRD were used as baselines. Meanwhile, convolutional neural network (CNN) method based on word embedding and Bi-LSTM method were chosen to compare disambiguation methods based on character and word levels, respectively.

To ensure the accuracy of experimental results, 10-fold crossing verification was adopted, and the statistics on final accuracy applied the mean of 10-fold-crossing-verified accuracy.

4.4 Experimental results and contrast analysis

In this experiment, Bi-LSTM neutral network model based on character level was applied, and 106 acronyms in MSH-WSD were used in the disambiguation test. The test results are listed in Table 1.

Table 1. Comparison of disambiguation results between Bi-LSTM method based on character-level features and non-neutral network method based on word-level features

Model methods	**AEC**	**MRD**	**2-MRD**	**UMLS SenseRelate**	**Character-level Bi-LSTM**
Accuracy	90.90%	87.59%	85.01%	83.00%	85.82%

Table 1 shows that the Bi-LSTM method based on character level shows a slightly lower accuracy than AEC and MRD, but it presents significantly higher accuracy than 2-MRD and SenseRelate method based on UMLS. Given that disambiguation methods based on character-level features do not require abundant corpus data to train word vectors, they are significantly more applicable than previous methods based on word level. The proposed disambiguation method achieves an ideal accuracy of 85.82% without using additional corpus data.

In the experiment, the proposed method was compared with the disambiguation methods based on word level that apply abundant external resources to train word embedding vectors, such as the CNN classification model of Kim [23] and Daojian [24], which was a CNN based on word level and supervised Bi-LSTM neutral network method based on word level. The results are listed in Table 2.

Table 2. Comparison of disambiguation results between neutral network method based on character-level features and neutral network method based on word-level features

Acronyms Model methods	Word-level CNN S200	Word-level Bi-LSTM S200	Character-level Bi-LSTM s200
Accuracy	94.91%	98.13%	85.82%

According to the comparative experiment, the accuracy of the disambiguation method based on supervised CNN, which is trained by word-level vectors, is 94.91%. The accuracy of the Bi-LSTM method based on word-level features is 98.13%, while that of the Bi-LSTM method based on character-level features is only 85.82%. Therefore, the neutral network method based on word level has a certain higher accuracy than the neutral network method based on character level. This finding is mainly attributed to the fact that the method based on word level uses mass external resources and thereby involves more effective external features. A detailed comparison is introduced in the

following text to disclose specific differences between the methods based on character-level features and those based on word-level features.

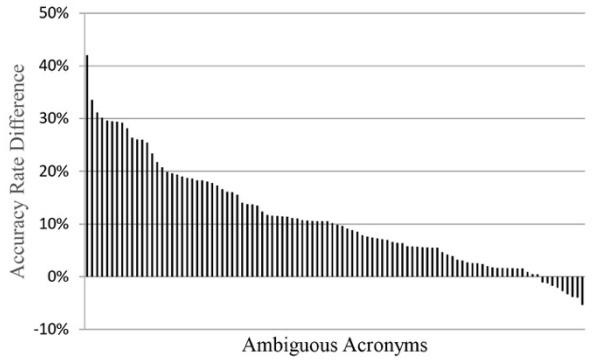

Fig. 7. Comparison between CNN model based on word-level features and Bi-LSTM model based on character-level features

Fig. 7 shows that for the acronyms in MSH-WSD, the CNN model based on word-level features is significantly superior to the Bi-LSTM model based on character-level features. In fig. 7, the horizontal coordinate refers to each ambiguous acronym and the vertical coordinate is the proportion of accuracy difference. The disambiguation accuracies of the acronyms are compared. The positive part reflects that the amplitude of accuracy of the CNN model based on word-level features is higher than that of the Bi-LSTM model based on character-level features, while the negative part reflects that the amplitude of accuracy of the CNN model based on word-level features is lower than that of the Bi-LSTM model based on character-level features. In Fig. 7, word-level feature model is significantly superior to character-level feature model and achieves higher accuracy in the disambiguation of acronyms, which is higher than 10% of the average accuracy. This result reflects that the disambiguation effect of the Bi-LSTM model based on character-level features is better than that of the CNN model based on word-level features with respect to acronyms in the MSH-WSD corpus.

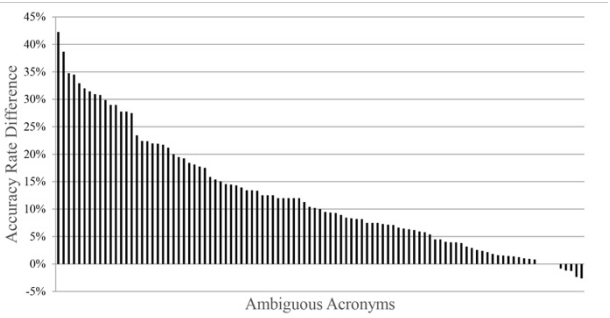

Fig. 8. Comparison between the Bi-LSTM model based on word-level features and the Bi-LSTM model based on character-level features

To perfectly compare and avoid disturbances of the model to the experimental results, the disambiguation methods of acronyms in MSH-WSD based on the Bi-LSTM model were compared. The comparison results are shown in Fig. 8. When the algorithm applies the Bi-LSTM in the late stage, the word-level feature model achieves higher accuracy than the character-level feature model.

The Bi-LSTM and CNN methods based on word-level features were compared with the Bi-LSTM method based on character-level features. The results are shown in Fig. 9. The blue line represents the methods based on character-level features, the black dotted line represents the CNN method

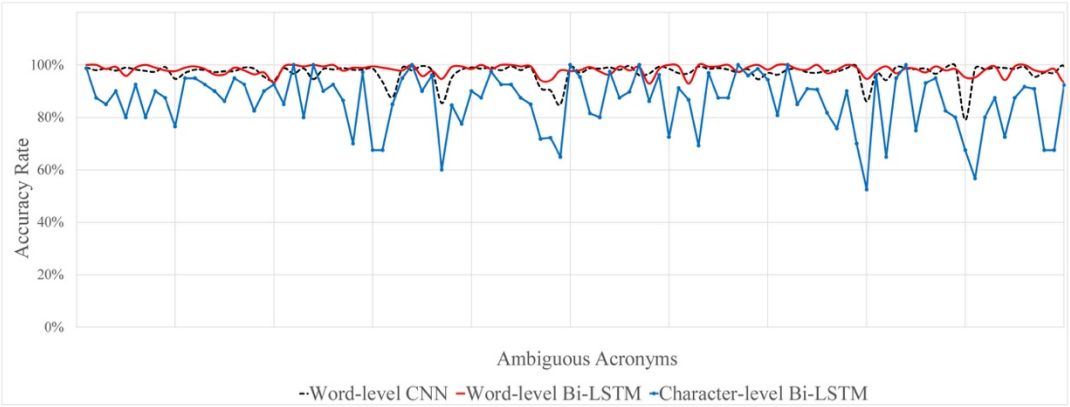

Fig. 9. Comparison between methods based on character-level features and the other two methods based on word-level features

based on word-level features, and the red solid line represents the Bi-LSTM method based on word-level features. Generally speaking, the Bi-LSTM method is the best. It is significantly superior to the CNN methods based on word-level features and is very stable. The disambiguation accuracy of the proposed method based on character-level features is also higher than 80% in most cases. Given that the methods based on character-level features do not require mass external resources to train vectors, they can obtain relatively ideal results. According to the observation results, the disambiguation methods based on character-level features are superior to two neutral network methods based on word-level features with respect to some acronyms. On this basis, the character-level vectors in some ambiguous acronyms can be deduced to contain features that are not available in word-level vectors. These useful features will be further explored in future studies to increase the disambiguation accuracy of biomedical acronyms.

5. Conclusions

In this study, a Bi-LSTM model based on character-level features is proposed for the disambiguation of biomedical

acronyms. The model achieves good performance. In addition, the Bi-LSTM model based on character-level features is compared with the traditional disambiguation model, the CNN model based on word-level features, and the Bi-LSTM model based on word-level features. The major conclusions drawn are as follows:

(1) Neutral network model based on character-level features needs no external biomedical resources and is superior to other non-neutral network methods in the disambiguation of biomedical acronyms.

(2) The accuracy of the neutral network model based on character-level features is at least 10% lower than that of the neutral network model based on word-level features because the neutral network model has certain advantages in feature extraction and expression with support from mass external training texts.

(3) Character-level feature model is compared with word-level feature model. The result shows that the latter has higher accuracy, which is attributed to the fact that in addition to the influence of external resources, the selected window size of training corpus in the character-level feature model is similar to that in the word-level feature model. In the same size of window, characters contain less context content than words, thereby missing some features and influencing the disambiguation effect of acronyms.

The Bi-LSTM model based on character-level features needs no external linguistic data and can still achieve satisfactory disambiguation of biomedical acronyms. It is applicable to the disambiguation of biomedical acronyms when the external resources are limited. Currently, character-level feature model cannot choose large window size owing to the limitations in computing ability. This limitation will influence the final disambiguation accuracy to some extent. To address this problem, future studies may focus on developing an appropriate method to compress character-level features and combine character-level and word-level features for the disambiguation of biomedical acronyms, with the aim of achieving remarkable disambiguation performance.

Acknowledgements

The authors are grateful for the support provided by the Fundamental Research Funds for the Central Universities, South-Central University for Nationalities (CZY18017) and the Natural Science Foundation of Hubei Province, China (Nos 2016CKC775).

References

1. Martinez, D., Baldwin, T., "Word sense disambiguation for event trigger word detection in biomedicine". *BMC bioinformatics*, 12(2), 2011, pp.S4-S4.
2. Kim, Y., Hurdle, J., Meystre, S.M., "Using UMLS lexical resources to disambiguate abbreviations in clinical text". In: *Proceedings of the 2011 International Conference on Annual Symposium proceedings*, Washington, USA: AMIA, 2011, pp.715-722.
3. Okazaki, N., Ananiadou, S., Tsujii, J., "Building a high-quality sense inventory for improved abbreviation disambiguation". *Bioinformatics*, 26(9), 2010, pp.46-53.
4. Sabbir, A., Jimeno-Yepes, A., Kavuluru, R., "Knowledge-based biomedical word sense disambiguation with neural concept embeddings". In: *Proceedings of the 17th International Conference on Bioinformatics and Bioengineering*, Virginia, USA: IEEE, 2017, pp.163-170.
5. Henry, S, Cuffy, C, McInnes, B., "Evaluating feature extraction methods for knowledge-based biomedical word sense disambiguation". In: *Proceedings of the 2017 International Conference on Biomedical Natural Language Processing*, Vancouver, Canada: ACL, 2017, pp.272-281.
6. Dongsuk, O., Kwon, S., Kim, K., "Word sense disambiguation based on word similarity calculation using word vector representation from a knowledge-based graph". In: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, USA: ACL, 2018, pp.2704-2714.
7. Duque, A., Stevenson, M., Martinez-Romo, J., "Co-occurrence graphs for word sense disambiguation in the biomedical domain". *Artificial intelligence in medicine*, 87, 2018, pp.9-19.
8. Jimeno-Yepes, A. J., 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, pp.223-236.
9. Jimeno-Yepes, A.J., McInnes, B.T., Aronson, A.R., "Collocation analysis for UMLS knowledge-based word sense disambiguation". *BMC Bioinformatics*, 12(3), 2011, pp.S4-S4.
10. Pasini, T., Navigli, R., "Train-o-matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: ACL, 2017, pp.78-88.
11. Panchenko, A., Marten, F., Ruppert, E., "Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: ACL, 2017, pp.91-96.
12. Charbonnier, J., Wartena, C., "Using Word Embeddings for Unsupervised Acronym Disambiguation". In: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, USA: ACL, 2018, pp.2610-2619.
13. Li, Y., Zhao, B., Fuxman, A., "Guess Me if You Can: Acronym Disambiguation for Enterprises". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia: ACL, 2018, pp.1308-1317.
14. REN, K., REN, Y.F., "Kernel Fuzzy C-Means Clustering for Word Sense Disambiguation in BioMedical Texts". *Journal of Digital Information Management*, 13(6), 2015, pp.411-420.
15. Yu, J., Jian, X., Xin, H., "Joint embeddings of chinese words, characters, and fine-grained subcharacter components". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: ACL, 2017, pp.286-291.
16. Chen, H., Huang, S., Chiang, D., "Combining Character and Word Information in Neural Machine Translation Using a Multi-Level Attention". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, New Orleans, USA: NAACL, 2018, pp.1284-1293.
17. Henry, S., Cuffy, C., McInnes, B., "Evaluating feature extraction methods for knowledge-based biomedical word sense disambiguation". In: *Proceedings of the 2017 International Conference on Biomedical Natural Language Processing*, Vancouver, Canada: ACL, 2017, pp.272-281.
18. Wang, L., Li, S., Sun, C., "One vs. many qa matching with both word-level and sentence-level attention network". In: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, USA: ACL, 2018, pp.2540-2550.
19. Liu, Q., Huang, H., Gao, Y., "Task-oriented word embedding for text classification". In: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, USA: ACL, 2018, pp.2023-2032.

20. Kumar, S., Jat, S., Saxena, K., "Zero-shot Word Sense Disambiguation using Sense Definition Embeddings". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: ACL, 2019, pp.5670-5681.
21. Raganato, A., Bovi, C.D., Navigli, R., "Neural sequence learning models for word sense disambiguation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: ACL, 2017, pp.1156-1167.
22. Le, M., Postma, M., Urbani, J., "A deep dive into word sense disambiguation with LSTM". In: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, USA: ACL, 2018, pp.354-365.
23. Kim, Y., "Convolutional neural networks for sentence classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar:ACL, 2014, pp.1746-1751.
24. Daojian, Z., Kang, L., Siwei, L., Guangyou, Z., Jun, Z., "Relation classification via convolutional deep neural network". In: *Proceedings of the 25th International Conference on Computational Linguistics*, Dublin, Ireland: ACL, 2014, pp.2335-2344.