Domain Neural Chinese Word Segmentation with Mutual Information and Entropy

Wang Jun
Science and Technology on Information Systems Engineering Key Laboratory, National University of Defense Technology
Changsha, Hunan, P.R. China
wangjun17c@nudt.edu.cn

Ge Bin
Science and Technology on Information Systems Engineering Key Laboratory, National University of Defense Technology
Changsha, Hunan, P.R. China
gebin@nudt.edu.cn

He Chunhui
Science and Technology on Information Systems Engineering Key Laboratory, National University of Defense Technology
Changsha, Hunan, P.R. China
xtuhch@163.com

ABSTRACT
Chinese word segmentation (CWS) is an important basic task for NLP. However, the word segmentation model trained by the generic domain corpus has a significant decline in performance in the word segmentation task oriented to the specific domain. Aiming at the features of domain segmentation, this paper using domain corpus as the training samples, and proposed combined with the terminology dictionary, new word detection and Bi-LSTM-CRF segmentation method to improve the problem of out-of-vocabulary (OOV). The word segmentation experiment was carried out on the corpus of the automotive domain. The results show that the precision and recall of the word segmentation have reached 0.95, and the value of F_1 also achieved 0.95, and they are better than state-of-the-art method. This method can also be combined with N-gram and chi-square statistic to further improve the recognition accuracy of OOV.

CCS Concepts
Computing methodologies→Artificial intelligence→Natural language processing→Lexical semantics

Keywords
Mutual Information; Entropy; Chinese Word Segmentation; OOV; Bi-LSTM-CRF

1. INTRODUCTION
In recent years, artificial intelligence has made breakthroughs in many key areas. The emergence of deep learning has driven the rapid development of NLP. As the most basic task of NLP, word segmentation directly affects the accuracy of subsequent part-of-speech tagging, named-entity recognition, and syntactic analysis. Unlike the basic unit of English being a word, Chinese has no obvious segmented mark, and the sentence appears in the form of a string. Therefore, the first step in dealing with Chinese is to segment, which is to convert 1-gram string into n-gram string.

ICIT 2019, December 20–23, 2019, Shanghai, China

ACM ISBN 978-1-4503-7663-1/19/12…$15.00
https://doi.org/10.1145/3377170.3377205

Chinese word segmentation can be divided into three types based on rules, statistics and deep learning. Rule-based segmentation is simple and efficient. When segmenting a sentence, each string of the statement is matched with the words in the vocabulary one by one, but maintaining the dictionary is a huge project, and has difficulty handling ambiguous words and OOV [1] . Statistic-based word segmentation needs to establish a statistical language model, divide phrases into words, and then calculate the probability of the results to obtain the segmentation method with the highest probability. Classic statistic-based segmentation includes Hidden Markov Model (HMM) and Conditional Random Field (CRF) [2-3]. The statistic-based segmentation does not require dictionary maintenance, and it can handle ambiguous words and OOV better. However, the effect heavily relies on the quality of the training corpus, and the calculation is much more than rule-based segmentation. In practical applications, we use both to improve the accuracy and domain adaptability. Deep learning brings new ideas to word segmentation. With the most basic vectorization features as input, through the multi-layer nonlinear transformation, the output layer can accurately predict the label corresponding to the current input[4]. Commonly used deep-learning networks include Convolutional Neural Networks (CNN), Long and Short Time Memory Networks (LSTM) [5].

As the word segmentation model trained by general corpus is not effective in the specific domain task，we propose a segmentation method with a domain corpus used as training sample, combined with a terminology dictionary, a new-word detection algorithm and Bi-LSTM-CRF model. In specific domain， this method can improve the recognition of terminology words and greatly improve the performance of OOV segmentation.

2. MODEL CONSTRUCTION
2.1 Sequence Annotation
The deep-learning model needs to learn the position features of each word. A label set is a standard for tagging word position information in text. At present, the application of Chinese word segmentation is more than 2-tag, 4-tag, and 6-tag label sets. In the deep-learning model for Chinese word segmentation, the most commonly use 4-tag label set (B, M, E, S). B represents the beginning, M represents the middle, E represents the end, and S represents a single word [6].Although the 6-tag label set is more comprehensive in place for the information expression of long words, the calculation amount is much larger. Considering this comprehensively, the paper uses the 4-tag label set to complete the annotation of the corpus.

2.2 Bi-LSTM-CRF Model

In the segmentation sequence annotation task, the current mainstream deep learning model is Bi-LSTM-CRF. The model uses a Bi-LSTM layer that can learn from both the forward and backward directions of the sentence. The output layer connects the CRF layer so that the model can acquire known features and potential features simultaneously and learn the state transition matrix. This makes it possible to determine the output labels associated with before and after. CRF quickly decodes the output of Bi-LSTM and the state transition matrix learned by CRF through the Viterbi algorithm to obtain the label sequence of the output [7]. This paper uses the architecture diagram shown in Figure 1 to design the Chinese word segmentation model:

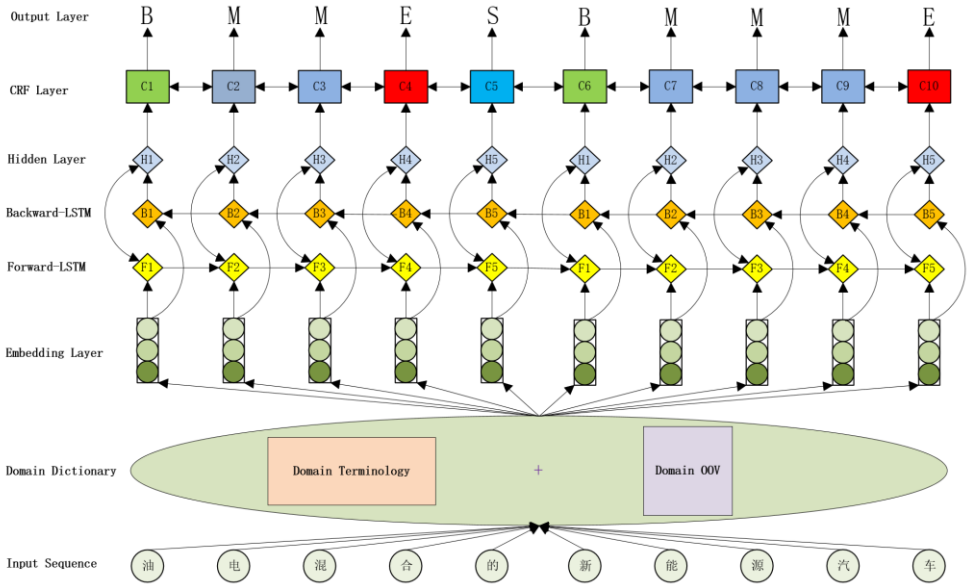

Figure 1. Bi-LSTM-CRF Chinese word segmentation model.

The first is the input layer, which mainly implements the input function of the corpus. Next is the domain dictionary, which contains two parts: domain terminology and domain new words. The domain dictionary is not one neural network layer in the actual sense. It is the feature that the segmentation model has learned from the domain dictionary to the domain words. Then there is the word embedding layer. The text first needs to complete the word vector embedding before inputting the word segmentation model. We use word2vec [8] and the dimension is set to 300 as the feature representation. The core part is the bidirectional long-term memory network layer, which uses Bi-LSTM to obtain context features. Then through the hidden layer the model achieves data conversion. Finally, the CRF layer is used to give the final sequence labeling result.

2.3 Domain New Word Detection Algorithm

2.3.1 Algorithm theory

To extract words from a piece of text, the criteria for word segmentation must first be clear. An idea occurs to us easily that counting word frequency. In the automotive field, for example, in the corpus, "of car" appeared 1358 times, while "car factory" only appeared 278 times, but we tend to regard "car factory" as a word. Because intuitively, the stability between "car" and "factory" seems to be higher. In order to calculate the stability of a text segment, it is necessary to enumerate the way it is combined. Mutual information (*MI*) is a concept in information theory that measures the degree of association between two signals. Pointwise mutual information (*PMI*) is a special case of mutual information used to calculate the degree of association between words in a text. We use *PMI* to measure the stability [9]:

$$PMI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

Attention only for stability of text segment is not enough, we need to see its external performance as a whole. Let's compare the Chinese binary words "AB" and "CD". If "AB" can be used like this: "ABD", "ABE", "FAB", "GHAB", etc., and "CD" only has "CDI" usage, which is very fixed, intuitively, we will not regard "CD" as a word since it can not be used freely. We can see that the text segment must have a certain degree of flexibility to form words. The more left and right word combinations, the more different contexts it will appear and the more likely it will form a word. We use information entropy to measure the flexibility of the left-neighbor word set and the right-neighbor word set of a text segment.

$$E_L(W) = -\sum_{\forall a \in A} P(aW|W)\log_2 P(aW|W) \tag{2}$$

$E_L(W)$ denotes the information entropy on the left of the text segment W. A is the set of characters appearing on the left of the text segment W. And $P(aW|W)$ denotes the conditional probability that the character on the left of the W is a. Similarly, the right information entropy of the text segment W can be expressed as:

$$E_R(W) = -\sum_{\forall a \in A} P(Wa|W)\log_2 P(Wa|W) \tag{3}$$

The flexibility of a text segment should be determined by the information entropy on both sides.We can find that E_R of "dis" and "neg" is much higher than E_L, and all kinds of nouns can be connected on their right side, while the left side is often a space, and neither of these two text fragments can be word-forming intuitively. Therefore, we directly define the flexibility of word-

forming of text segments as the lower value of E_L and E_R. The stability and flexibility are indispensable to the standard of word-forming of text segments. Mutual information alone can extract a lot of semi-finished products like "Vol", "lvo". Information entropy alone can extract a large number of infixes, which is more obvious in the domain text, such as "pe" in "piperidine", "pipecoline" and "lupetidine" in the domain of chemistry, which represents complete oxidation. In order to improve the accuracy, we set a *PMI* threshold of 20 to exclude some segments with low stability. We define *Score* as an evaluation index for the possibility of word-forming. The higher the score, the greater the possibility of word-forming.

$$Score = PMI + \min\left[E_L(W), E_R(W)\right] \quad (4)$$

2.3.2 Algorithm flow

We first use precise mode of jieba segmentation[10] to make a rough segmentation with original domain corpus. Chinese word segmentation needs to read the lexicon repeatedly. In order to minimize the unnecessary segment comparison and improve the query efficiency, we use the Trie [11], which is the most suitable data structure for storing the dictionary and calculating the word frequency. In order to calculate E_L and E_R, we need to combine the segment with the prefix and suffix. Therefore, we need to store the segment with length 3. We use the tri-gram method to build the node, and use the Trie to store the word segmentation, such as: [fuel, cell, system] → [fuel, cell, system, fuel cell, cell system, fuel cell system]. In single text we use Trie to calculate *PMI*, E_L and E_R of the text segment and get Score. We sort by *Score* and take top 3 as new words. Figure 2 is the flow chart of the new word detection algorithm.

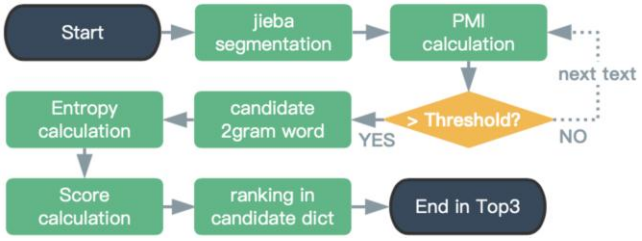

Figure 2. New word detection algorithm flow chart.

3. REALIZATION THEORY OF DOMAIN CWS SYSTEM

3.1 Domain Language Resource Obtaining

The word segmentation model trained by public dataset has a low accuracy rate for the segmentation of specific domain corpus, because its domain knowledge learned from the corpus in the process of training is small. If the model can learn from corpus with rich domain knowledge, its generalization performance will be greatly improved.

Language resources are derived from the perspective of NLP applications, and are generally classified into two types: domain corpus and domain dictionary.

The domain corpus is from the perspective of the corpus, and the corpus here is defined as text level (a collection of texts formed at the level of natural sentences, that is, sentences, paragraphs, chapters, etc.). The domain dictionary refers to a language resource library formed by sentence-level language units. The language units of this level can be strokes, radicals, words, phrases, and so on. This paper takes the automotive field as an example and gives the specific process of corpus and dictionary construction.

3.1.1 Domain corpus construction

The corpus resources in the automotive field come from NetEase[1]. In addition to a large number of domain terminology, the news texts in the automotive field also contain new words and hot words in the field. The refinement of news text ensures that the text expands around several keywords, therefore can improve the performance of the new word detection system. We use web crawlers to get all the news links through regular expressions, then crawl and store the corresponding news in txt format, for a total of 4,000 documents, to complete the small-scale auto domain corpus construction task.

3.1.2 Domain dictionary construction

The terminology dictionary mentioned here mainly refers to the domain feature dictionary, which is a collection of words that are strongly related to the field and are able to distinguish between fields, for example, "BMW" and "Chevrolet". These words are often used as classification features. The construction methods of domain dictionary can be divided into: artificial building and maintaining domain dictionary; combining corpus and statistical methods to construct domain dictionary; using Wiki mining to build domain dictionary [12].

In view of the large amount of computation based on Wiki link construction and high requirements on computing equipment, this paper uses the terminology dictionary combined with corpus and statistical methods (new word detection algorithm based on mutual information and entropy) to construct the lexicon. This method can detect new domain terminology while ensuring the quality of the dictionary.

The terminology dictionary in this paper is mainly obtained through the dictionary published by Sogou[2], a total of 3,734 words. In addition, 2000 texts were randomly selected from the domain corpus, and the candidate words detected by algorithm in each text are only extracted top 3, totaling 6,000 candidate words. It was found that taking the candidate words of top 3 can better mine the key information of the text while ensuring the accuracy. All candidate words are deduplicated and denoised, and finally 3005 valid new words are obtained. Next the domain terminology dictionary is deduplicated and merged, and ultimately a domain dictionary containing 6477 automotive fields is obtained.

3.2 Dataset Division

When training a supervised machine learning model, the dataset is divided into train, validation, and test datatest. The original corpus is divided in order to be able to select the best generalization model and prevent the model from overfitting. The experiment divided 4,000 articles of automotive news by 0.8:0.15:0.05 (train dataset: validation dataset: test dataset). The train, validation, and test datasets are preprocessed by sequence labeling. The train and validation datasets are used to complete the training and validation of the deep neural network model, and the test dataset is used to evaluate the performance of the model.

3.3 Realization Steps

Dictionary-based word segmentation does not handle a large number of OOV in domain corpus, and the quality of segmentation based on statistics needs to be improved. The domain neural Chinese word segmentation method based on mutual information and entropy proposed in this paper combines external dictionary and statistical algorithm. Under the premise of guaranteeing the

[1] https://auto.163.com

[2] https://pinyin.sogou.com/dict

quality of the dictionary, the unsupervised method can be used to extract more potential OOV from the domain corpus. Figure 3 is a flow chart of our model. The specific steps are as follows:

(1) Obtaining automotive terminology dictionary and automotive news from the Internet as domain corpus.

(2) Using the new word detection algorithm based on mutual information and entropy to mine the domain corpus automatically, generating a new word dictionary in the automotive domain, and merging with the terminology dictionary to obtain a high-quality domain dictionary containing plenty of new words.

(3) Using the jieba word segmentation module, adding the merged domain dictionary as a custom dictionary, and accurately segmenting the corpus, finally obtaining the annotated domain corpus generated by adding the merged domain dictionary.

(4) Using the annotated domain corpus generated by adding the merged domain dictionary to train the Bi-LSTM-CRF deep learning model to finish the domain Chinese word segmentation.

(5) Using the domain CWS model obtained to segment the test corpus and obtain the final word segmentation result.

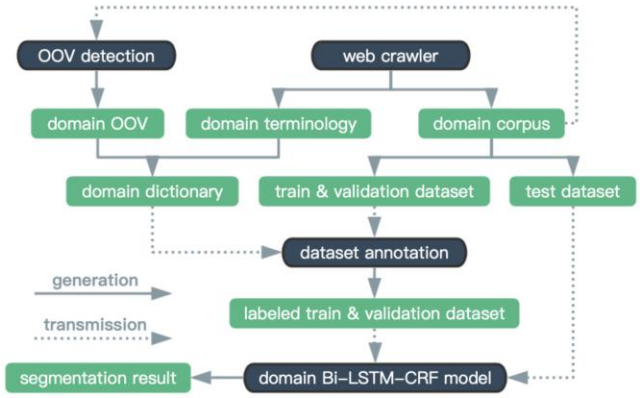

Figure 3. Neural Chinese segmentation model flow chart.

4. EXPERIMENT DESIGN AND ANALYSIS

4.1 Experiment Design

The domain test dataset concludes 200 original text, a total of 185,371 words. The word segmentation experiment is divided into six groups: Experiment 1 used the People's Daily corpus published by Peking University to train the model, and conducted word segmentation on the original test dataset of pku[3]; Experiment 2 used pku corpus to train the model, and conducted word segmentation on the domain test dataset; Experiment 3 used the public corpus provided by Microsoft Research Asia to train the model, and conducted word segmentation on the original test dataset of msr; Experiment 4 used msr corpus to train the model, and conducted word segmentation on the the domain test dataset; Experiment 5 used the domain corpus to train the model, and conducted word segmentation on the domain test dataset; Experiment 6 used the domain corpus generated by adding the domain merged dictionary to train, and conducted word segmentation on the domain test dataset.

4.2 Experimental Results and Analysis

4.2.1 Hyper-parameters setting

The hyper-parameter settings used by Bi-LSTM-CRF are shown in Table 1.

[3]http://sighan.cs.uchicago.edu/bakeoff2005/

Table 1. Core Hyper-parameters setting

Parameter name	Value
Number of hidden layers	300
Learning rate	$1e^{-3}$
Embedding dimension	300
Dropout rate	0.5
Batch size	32
Epoch	30
Optimizer	Adam

4.2.2 Experimental results and analysis

In order to compare the effects of Chinese word segmentation in different experiments, we use the *Precision*, *Recall* and F_1 value to evaluate the performance of the model. Table 2 gives the results of the experiment, and Table 3 lists the results of the word segmentation of a sample passage in the test corpus in experiment 2, 4, 5, and 6. The sample passage given in Table 3 means "Toyota's Rand Cool Road and Prado are good at off-road weapons, and have a good reputation. // BYD is waiting for the right time to promote the commercialization of new energy vehicles in the domestic and foreign markets.// ML3004MATIC equipped with Mercedes-Benz 4MATIC full-time four-wheel drive system, the gasoline engine can emit 300 Nm of peak torque in a very short time." From the results in Table 2, the performance of the model trained with pku and msr public general corpus on the test dataset in automotive domain is far less than the performance on the original universal test dataset, showing the poor generalization ability. Compared with the relatively good pku corpus, the model trained with domain corpus improves the *Precision*, *Recall*, and F_1 value by 12%, 11%, and 11%, respectively. On this basis, combined with the domain merged dictionary, the *Precision*, *Recall*, and F_1 value can be increased by 15%, 17%, and 16 % respectively. From the experimental results, in the Chinese text segmentation task of the automotive field, the domain neural Chinese word segmentation method based on mutual information and entropy can greatly improve the accuracy and generalization ability of the model after combining the domain dictionary and domain corpus.

Table 2. Results of Bi-LSTM-CRF Test on Different Datasets

Datasets	*Precision*	*Recall*	F_1
Exp2 pku.train+auto.test	0.80	0.78	0.79
Exp3 msr.train+msr.test	0.97	0.97	0.97
Exp4 msr.train+auto.test	0.77	0.77	0.77
Exp5 auto.train+auto.test	0.92	0.89	0.90
Exp6 auto.train+dict+auto.test	**0.95**	**0.95**	**0.95**

5. CONCLUSION AND FUTURE WORK

On the Chinese word segmentation task, the word segmentation model trained on public dataset is not able to segment the domain corpus very well. We propose a domain neural Chinese word segmentation method based on mutual information and information entropy. The results show the method can effectively improve the word segmentation accuracy and generalization ability of the word segmentation model in specific domain corpus. Next, we will try to

extract the part-of-speech(POS) features of the words from domain corpus and integrate them into the Bi-LSTM-CRF word segmentation model to further improve the performance of the word segmentation algorithm.

Table 3. Word segmentation results On automative test sample with Bi-LSTM-CRF model trained by different datasets

Experiment	Segmentation Results of Sample
Experiment 2	丰田 旗下 的 兰德 酷路 泽及 普拉多 都 是 擅长 越野 的 利器， 有着 不错 的 口碑 。//比亚迪 正在 等待 合适 时机 推动 新 能源 汽车 于 国内外 市场 的 商业化 普及 进程 。//ML 3004MATIC 搭 载 了 奔驰 4MATIC 全时 四 驱 系统， 汽油 发动机 能 在 极 短 时间 内进 发出 300 牛米 的 峰值 扭矩 。
Experiment 4	丰 田旗 下 的 兰德酷路泽 及 普拉 多 都是 擅长 越野 的 利器， 有着 不错 的 口碑 。//比亚迪 正在 等待 合适 时机 推动 新 能源 汽车 于 国内外 市场 的 商业化 普及 进程 。//ML 3004MATIC 搭载 了 奔驰 4M AT IC 全时 四驱 系统， 汽油 发动机 能 在 极 短 时间 内 迸发 出 300 牛米 的 峰值 扭矩 。
Experiment 5	丰田 旗下 的 兰德 酷路泽及 普拉多 都 是 擅长 越野 的 利器， 有着 不错 的 口碑 。//比亚迪 正在 等待 合适 时机 推动 新能源 汽车 于 国内外 市场 的 商业化 普及 进程 。//ML3004MATIC 搭载 了 奔驰 4 M ATIC 全时 四驱 系统， 汽油 发动机 能 在 极短 时间 内 迸发出 300 牛米 的 峰值 扭矩 。
Experiment 6	丰田 旗下 的 兰德酷路泽 及 普拉多 都是 擅长 越野 的 利器， 有着 不错 的 口碑 。//比亚迪 正在 等待 合适 时机 推动 新能源汽车 于 国内外 市场 的 商业化 普及 进程 。//ML3004MATIC 搭载 了 奔驰 4MATIC 全时 四驱 系统， 汽油发动机 能 在 极短 时间 内 迸发 出 300 牛米 的 峰值扭矩 。

6. ACKNOWLEDGMENT

This research is supported by the NNSF of China under No.61872446 and No.61902417.

REFERENCES

[1] Zaharia, M. and Chowdhury, S. 2010. Cluster Computing with Working Sets. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*.

[2] Zhang, H. P., Yu, H. K., Xiong, D. Y. and Liu, Q. 2003, July. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the second SIGHAN workshop on Chinese language processing*-Volume 17 (pp. 184-187). Association for Computational Linguistics.

[3] Murdoch S.J., Lewis S. 2005. Embedding Covert Channels into TCP/IP. In: Barni M., Herrera-Joancomartí J., Katzenbeisser S., Pérez-González F. (eds) Information Hiding. IH 2005. *Lecture Notes in Computer Science*, vol 3727. Springer, Berlin, Heidelberg. DOI=https://doi.org/10.1007/11558859_19

[4] Du X, Cai Y, et al. 2016. Overview of deep learning. 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC). IEEE, 159-164.

[5] Wang, P., Qian, Y., Soong, F. K., He, L., and Zhao, H. 2015. A unified tagging solution: Bidirectional lstm recurrent neural network with word embedding. *arXiv preprint arXiv:1511.00215*.

[6] Ma, J., Ganchev, K., and Weiss, D. (2018). State-of-the-art Chinese word segmentation with bi-lstms. *arXiv preprint arXiv:1808.06511*.

[7] Jin, Y., Xie, J., Guo, W., Luo, C., Wu, D., & Wang, R. 2019. LSTM-CRF Neural Network with Gated Self Attention for Chinese NER. IEEE Access.136694-136703.

[8] Caselles-Dupré, H., Lesaint, F., and Royo-Letelier, J. 2018, September. Word2vec applied to recommendation: Hyperparameters matter. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 352-356). ACM.

[9] Xiaojuan, F. H. K. S. Z., and Wenbiao, X. 2005. Chinese Word Segmentation Research Based on Statistic the Frequency of the Word. In Computer Engineering and Applications.

[10] Sun, J. 2012. 'Jieba'Chinese word segmentation tool. https://github.com/fxsjy/jieba.

[11] Liu, J., Wu, F., Wu, C., Huang, Y., and Xie, X. 2019. Neural Chinese word segmentation with dictionary. Neurocomputing, 338, 46-54.

[12] Yin, W., Zhu, M., and Chen, T. 2013, July. Domain Thesaurus Construction from Wikipedia. In International Conference on Computer, Networks and Communication Engineering (ICCNCE 2013). Atlantis Press.