Detecting and Finding the True Meaning of Jargons

Yingying Lao
Graduate School of Integrated Basic Sciences, Nihon University, Tokyo, Japan

Chenghuan Zhang
Graduate School of Engineering, The University of Tokyo, Tokyo, Japan

Yilun Wei
Graduate School of Integrated Basic Sciences, Nihon University, Tokyo, Japan

Dongli Han
Department of Information Science, Nihon University, Tokyo, Japan
han@chs.nihon-u.ac.jp

ABSTRACT

While looking at the online bulletin boards, some obscure description could be found in the threads. In most cases, it is usually because that the authors are not willing to mention the true name of somebody or something directly, and in hence take the place of it with a substitute expression which has the same attributes, such as homophonic word or objects which appear similarly. These words and phrases, which are used irregularly in the context, could be called jargons. In this study, we propose a method to detect and understand jargons by considering the relevance of context words in word-embedding expressions. Specifically, we first calculate the topic correlation-coefficient of each word pair occurring in the text to find the jargon. Then, we combine a set of contextual information to find out a list of candidate words which probably deliver the true meaning of the jargon. By comparing the candidate words and the jargon word from 3 aspects (meaning, pronunciation and shape), the true meaning of the jargon could be determined based on the aggregative similarity. Finally, we have conducted an experiment to examine the usefulness of our approach.

Additional Keywords and Phrases: Jargon, Word-embedding expression, Word-vectors, Similarity, Natural processing language.

CCS CONCEPTS

• **Information systems**; • **Information retrieval**; • **Retrieval tasks and goals**; • **Information extraction**;

ACM Reference Format:
Yingying Lao, Chenghuan Zhang, Yilun Wei, and Dongli Han. 2021. Detecting and Finding the True Meaning of Jargons. In *2021 The 7th International Conference on Frontiers of Educational Technologies (ICFET 2021), June 04–07, 2021, Bangkok, Thailand.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3473141.3473225

1 INTRODUCTION

In general, a jargon is defined as the usage of a specific phrase or word in a particular situation. These specialized terms are used to convey hidden meanings accepted and understood in that field. It works to provide a technical language to meet the needs of the group of people working within the occupation and to improve in-group efficiency. For instance, Hotel is one of the job-fields that use jargon as a communication tool [1]. In front office department of hotels, words such as 'Queen', 'CO' are usually used, which have different meanings than usual. In this context, 'Queen' presents a type of bed which is big enough for two people, and 'CO' is an abbreviation that stands for 'Check Out', which means that a guest has returned the room key to front office. It has been found that the forms of jargons used can be classified as abbreviation, acronyms, word and phrase formation, and it is unintelligible for other people who do not belong to that particular profession.

However, jargon also could be found in social intercourse. By analyzing the conversation and dialogue on WhatsApp, Dewi et al have found that the use of language chat among lecturers in the Islamic State University of North Sumatra is considered the jargon Indonesia, English, and mixed between Indonesia and English [2]. On the other hand, information exchange of illegal and harmful activities on BBS (Bulletin Board Systems) through jargon expressions make it difficult to extract information using standard tools [3]. These jargons are invented on purpose with the abbreviation, acronyms, and phrases in that specific situation. Even though the meaning of the formed term is not the original meaning or usual usage of the word, people in that particular group could understand it easily than others.

There have been several previous studies concerning the detection of jargons used in different circumstances based on Natural Language Processing (NLP) techniques. For jargons delivering harmful information, Abiko et al. contributed a database of jargon which contained the commonly used jargon examples from three BBS, and used it to normalize the expression of text automatically to filter harmful information containing jargons on the BBS [3]. Ohnishi et al also proposed a method based on word appearance distribution, using a known database of jargon for collecting the new jargon cases from underground BBS [4]. On the other hand, jargons used irregularly in social media also deserved investigation. Aoki et al. focus on non-standard usage of common words on Twitter [5-6]. They converted each word into a vector using word-embedding expression, and detected the potential jargon by calculating similarity of vectors between the target word and the context word in different corpora. Their study shows a view of point which is similar to ours to find out the jargon by measuring the relevance among context words.

In the above-mentioned studies, jargons have been discovered in some way based on their contexts, but the meaning of the word

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICFET 2021, June 04–07, 2021, Bangkok, Thailand
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8972-3/21/06…$15.00
https://doi.org/10.1145/3473141.3473225

still remains confusing for readers that are not familiar with the particular field. We therefore hope to propose a method to detect and understand jargons by considering the relevance of context words in word-embedding expressions. Word-embedding expression is a type of word representation that has been trained from a large number of text and allows words with similar meaning to have a similar representation (word vector). Taking advantage of this merit, we consider that it is possible to generate a word vector based on context information, and compute the similarity between the generated vector and the jargon word vector to determine the true meaning of a jargon [7-9]. Specifically, we suppose that a piece of text usually describes a theme, and each word in the context is in the service to present the topic. We utilize word-embedding expressions to determine and calculate the topic correlation-coefficient of each word in the text. When the topic coherence-coefficient of a word is lower than the specified threshold, it means that the jargon might exist and is found in the similar way. Then, we take into account of the contextual information and find out a list of candidate words. By comparing candidate words and the jargon in 3 aspects (meaning, pronunciation and shape), we can rank the candidate words and expect the most similar one to appear at the top of the rank. Finally, we have conducted an experiment to examine the usefulness of our approach.

This paper is structured as follows. We first describe our approach for detecting and finding true meaning of jargons by considering the relevance of context words in details in Section 2. Then in Section 3, we show the results of experiments utilizing our approach. In the end, a summarization of this study and future works will be provided.

2 OUR APPROACH

Here in this paper, we focus on jargon expressions in Japanese. Our goal is to detect and understand jargons by considering the relevance of context words in word-embedding expressions.

In this section, we first describe a customized method for detecting the jargon based on contextual words. Next, we elaborate how to analyze the meaning of jargons by combining various contextual information with a set of candidate words which probably deliver the similar meaning of the jargon. By comparing the candidate word and the jargon word from 3 aspects (meaning, pronunciation and shape), the true meaning of the jargon could be determined based on the aggregative similarity.

2.1 Jargon Detection

As mentioned above, jargon is not the original meaning or usual usage of the word, and is not easily distinguished except people who are familiar to the field. In this study, we pursue a method to detect the jargon expression in a text only by considering the contextual information. As a preliminary investigation, we have collected 200 reviews from Japan Apple Store and converted each word in the text into a vector with word-embedding expression in advance [10]. Reviews posted in Apple Store, such as assessments of applications, are commonly composed of several sentences and focused on a certain topic, and all words appearing in the text are usually closely relevant to the topic. In other words, if an irregularly used word appears in the review, its relevance to the original topic of the text will be reduced. To judge whether the irregularly used word (target jargon) exists or not in a text, we incorporate the topic coherence-coefficient score to measure the relevance of all words in a text to the topic, as shown in Formula (1) and (2).

$$AP_{s,t} = -\left(\sum_{w \in s} R_{w,t} \log\left(R_{w,t}\right)\right) \quad (1)$$

$$C_s = \max\left(AP_{s,t}\right) \quad (t \in T = 200) \quad (2)$$

Given a text s, we focus on the topic coherence between the text and the topic, and calculate the topic coherence score with the above formulas. In our preliminary work, each word in the review will be converted into a 200-dimensional vector by word embedding, and the value of each dimension represents the association of the word with a topic in the review collection. In Formula (1), $R_{w,t}$ represents the association of the word w and the topic t, and $AP_{s,t}$ indicates the topic coherence between the text s and the topic t. As there are 200 topics in our data collection, the topic coherence score C_s is determined by the maximum $AP_{s,t}$. When the C_s gets lower than the threshold which is computed in our preliminary experiment, it implies that at least one irregularly used word (target word) is likely to exist in the text. After confirming that the target word exists in the text, the next step is to determine which word it is. Here, we employ the method Aoki et al. proposed [5] for measuring this by calculating the similarity between each word with Formula (3).

$$score_{w_t} = \frac{\sum_{w_j \in w_c} \sin\left(v_{w_t}\ ,\ v'_{w_j}\right) \times \alpha_{w_j}}{\sum_{w_j \in w_c} \alpha_{w_j}} \left(0 < score_{w_t} \leq 1\right) \quad (3)$$

In Formula (3), $score_{w_t}$ represents the irregularly used degree of the t^{th} word. This is computed by the t^{th} word and its surrounding words in word-embedding expressions. $\sin(v_{w_t}\ ,\ v'_{w_j})$ represents the similarity between the t^{th} word and the surrounding j^{th} word (c indicates the surrounding scope of calculation), and α_{w_j} stands for the weight of the surrounding j^{th} word calculated by its distance to the t^{th} word. When $score_{w_t}$ gets lower than the threshold that we have set up in advance, we consider that the word is likely to be a jargon.

2.2 Jargon Meaning Inferring

In this section, we elaborate the method to infer the true meaning of the jargon by recommending similar words. These similar words are selected by ranking the aggregative similarity to the jargon word which is computed based on the contextual information. As Figure 1 shows, we find candidate words and calculate the aggregative similarity in two steps. Firstly, we find a list of candidate words that are considered to be approximate to the jargon used in the text. Secondly, we calculate the aggregative similarity of the candidate word and the jargon. The aggregative similarity of a candidate word is computed from 3 aspects: meaning, pronunciation and shape. Finally, we rank the aggregative similarity of candidate words, and hope that the most similar word would appear at the top of the candidate list. In the following parts, we describe each step in details.

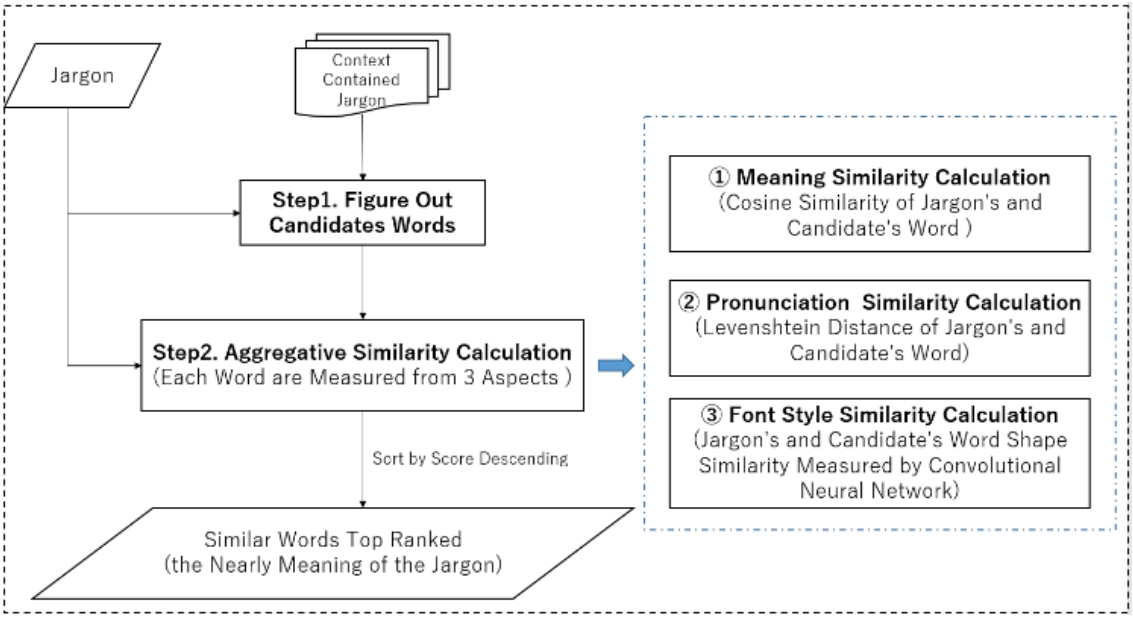

Figure 1: The Process of Inferring the Meaning of Jargon

2.2.1 Finding the Candidate Words. Here, we consider that analyzing the usage of the jargon in the corpus is important. Taking into account the jargon and other words appearing simultaneously in different cases, it may help us to infer the true meaning of the jargon word. Therefore, we designed a neural network model to condense the information of the jargon usage in the corpus. For a jargon, four kinds of information are collected.

- 1) the title of the thread which contains the jargon;
- 2) the full text of the thread which contains the jargon;
- 3) the sentence containing the jargon;
- 4) words around the jargon.

The above textual data will be converted into a 200-demenssion vector through the neural network model. Then, through calculating and sorting the Cosine Similarity between the converted vector and each word vector in the data collection, we could acquire a list of candidate words that may be relevant to the jargon. It is conducive to narrow down the workload of jargon semantic analysis in this way.

2.2.2 Aggregative Similarity Calculation. As shown in Figure1, to find the true meaning of the jargon in the text, we have measured the aggregative similarity to rank the candidate words selected in section 2.2.1, and hope the most similar word will appear at the top of the ranked list. Specifically, we compare the jargon and each word of the candidate words from 3 aspects.

- 1) Meaning Similarity: we calculate the Cosine Similarity of the Jargon's and the Candidate Word's vector based on word-embedding expression. When the Cosine Similarity value is close to 1, it implies that the meaning of candidate word is close to the jargon used in the text.
- 2) Pronunciation Similarity: As Japanese pronunciation can be expressed in Roman letters, we transform the jargon word and the candidate word to the Roman-letter form, and compute the editing distance of two Roman words by Levenshtein distance. Similarly, when the pronunciation similarity value is close to 1, it implies that the pronunciation of candidate word is close to the jargon used in the text.
- 3) Shape (font-style) Similarity: Some jargons in Japanese might be expressed as the division, combination or partial extraction of characters. For example, the character '魏' could be divided into two characters '委' and '鬼'. In order to distinguish these irregular expressions in Japanese, we have referred to a previous study [11], and used 12 different Japanese fonts for training a convolutional neural network model (CNN model) as shown in Figure 2. With the CNN model, the shape (font style) similarity between the jargon and the candidate word can be calculated and represented as a score. When the score is close to 1, it implies that the shape of candidate word is close to the jargon used in the text.

With the above-mentioned three similarity calculations, each candidate word will get three scores, and the aggregative similarity takes the maximum score. Finally, we can rank the candidate words by the aggregative similarity. Some examples with the proposed method will be demonstrated in Section 3.2.

3 EVALUATION EXPERIMENTS

In this section, we show the evaluation experiments and their results for examining the effectiveness of the proposed method. Firstly, an evaluation experiment for verifying the usefulness of our jargon detection approach will be described in Section 3.1. Then, we describe the evaluation experiment and show some examples for our approach to infer the true meaning of jargons in Section 3.2.

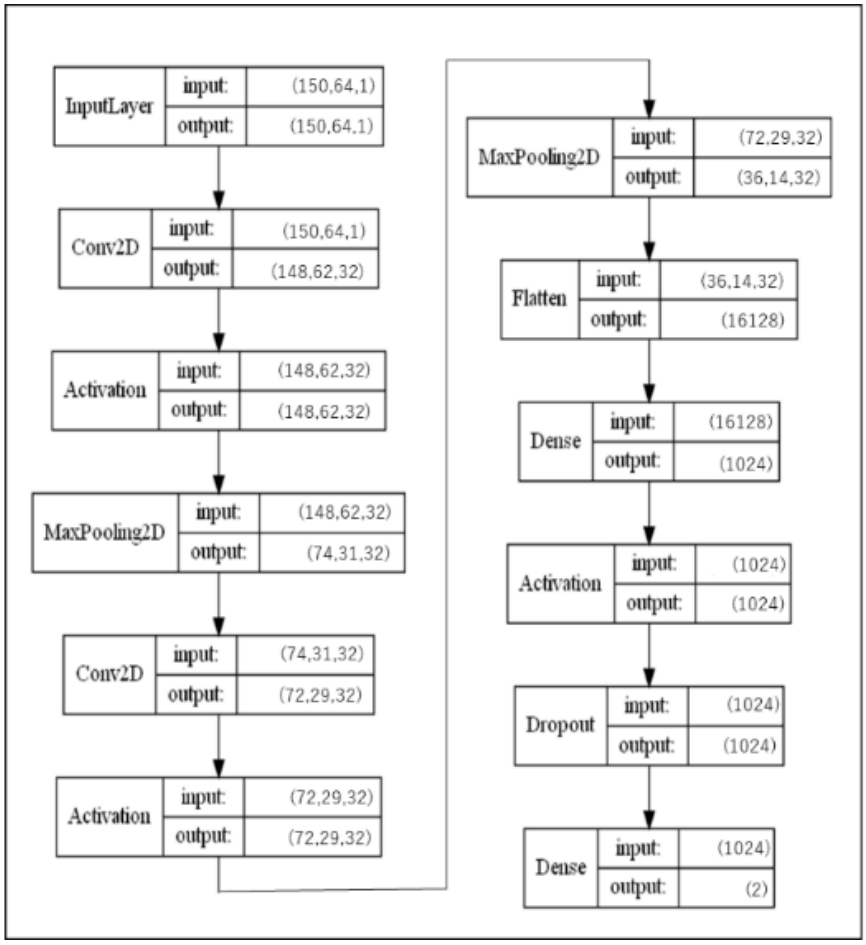

Figure 2: The Process of Calculating the Shape Similarity between Two Expressions based on a CNN Model

3.1 Jargon Detection Evaluation

The jargon detection approach we have described in Section 2.1 consists of two parts: 1) to determine whether the jargon word exists or not in a text, and 2) to find the jargon word.

Therefore, we have conducted two experiments for evaluating the effectiveness of each part, in Section 3.1.1 and Section 3.1.2.

3.1.1 Locating the Text. Here, for examining the usefulness of our approach to locate the sentence containing the jargon, we have used the 200 reviews extracted from Japan Apple Store to generate three test data-sets as shown below:

- 1) 200 reviews are randomly divided into dataset 1 and 2. Each dataset contains 100 reviews.
- 2) For each review in dataset 1, a word is selected randomly and replaced with another word which has the same pronunciation but different meaning. All the replaced reviews compose the dataset 3.

With dataset 1, 2 and 3, we calculate the topic coherence-score of each text with the method introduced in Section 2.1. Besides, as the baseline method, we used 'arithmetic mean' to calculate the topic coherence between the text and each topic. The results are shown in Table 1. Through the comparison between the results of dataset 1 and 3, we find that the mean topic coherence-score of each dataset drops down due to the replaced jargons in the text. Dataset 2 which contains different reviews also shows the similar tendency.

Then, to examine the accuracy for detecting jargons with above methods, we collected 50 texts containing jargon expressions from Japanese BBS '2ch'[1], and 100 texts without jargons from Japan Wikipedia as the test dataset 4, and carry out a same comparison experiment. Through the results shown in Table 2, It is clear that our proposed method performs better than the baseline one, where the F score reached 0.901.

3.1.2 Finding the Jargon. In our study, we applied Aoki et al.'s method to measure and extract the jargon used in a text by calculating the similarity between each context word. To evaluate the extraction of 150 texts (100 texts replaced jargon randomly in dataset 3 and 50 texts containing jargon in dataset 4), its accuracy reaches 0.768.

[1] https://2ch.sc/

Table 1: Evaluation on the usefulness of jargon detection

Method	The Mean Topic Coherence Score of Dataset		
	Dataset 1	Dataset 2	Dataset 3
Baseline Method	3.9706	3.9195	3.8501
Proposed Method	3.6971	3.6930	3.6862

Table 2: Evaluation on the effectiveness of jargon detection

Baseline Method	Predicted as Texts Containing Jargons	Predicted as Texts without Jargons
Texts Containing Jargons	50	0
Texts without Jargons	100	0
Proposed Method	Predicted as Texts Containing Jargons	Predicted as Texts without Jargons
Texts Containing Jargons	44	6
Texts without Jargons	1	99

Table 3: Ranking results by proposed method (two patterns)

	The Average Position of the Word with Same Meaning as the Jargon	Average Rank of the Words with Similar Meanings to the Jargon (Similarity> 0.5)	Average Rank of the Words with Similar Meanings to the Jargon (Similarity> 0.6)
Pattern 1	4,625	2,186	2,605
Pattern 2	2,748	2,145	2,518

3.2 Jargon Meaning Inferring Evaluation

In Section 2.2, we have proposed a method to find candidate words and calculate aggregative similarity of each word. By ranking candidate words with the aggregative similarity score, we expect that those words with similar meanings to the jargon can be ranked at the top. With the aggregative similarity score, we consider two patterns to rank:

- 1) the aggregative similarity score;
- 2) the aggregative similarity score multiplied by the cosine similarity of candidate word vector and mean vector of surrounding words.

Here, we have collected 800 texts with the total number of words to be 34,895 containing jargons from Japanese BBS '2ch' and employed our approach to verify the effectiveness of our approach. Here, we have used the above 2 patterns to rank the candidate words selected in each case, and the average ranking results are shown in Table 3. As we can see here, although the true word meaning of the jargons are not able to appear in top 10, or even top 100, we can still decrease the searching times from 34,895 to 2,100 ~ 4,600. It shows the usefulness of our approach in a sense. Moreover, by adjusting the threshold values of the two word's similarity could narrow down the search scope much more.

At the end, we would like to show two examples. Example 1 is the text "一年乙 来年は分かりやすい相だといいな". The jargon word "乙" used here means "Finish a Good Work", instead of its usual meaning "Celestial Stems and Branches". As we calculate for each candidate word, the word "わら" meaning "finish" is found to be ranked 2,079th in the candidate word list, which is far from what we have imagined.

In another successful example, the jargon word "オワタ", originally meaning "Sad" or "Broken Heart", is used to express the meaning of "finish" or "accomplish something". Luckily this time, we are able to find the word "わら" ranked 12th in the candidate word list which is considered to have the similar meaning as the jargon.

4 CONCLUSION

In this study, we have proposed a new method for detecting and finding true meaning for the jargon used in a text by considering the relevance of context words in word-embedding expressions. Our approach is applied to detect the jargon expression from Japanese BBS '2ch', and is proved to be useful in locating the text containing the jargon word, and reducing workload during the process of finding the true meaning for the jargon expression.

In future studies, we are about to improve the calculation of aggregative similarity to find out the word meanings approximate to the jargon, and develop a jargon analyzing system based on the idea we have raised in this paper.

REFERENCES

[1] Komang Tia Dwi Pradipta ., Dr. I Gede Budasi, M.Ed.., Putu Eka Dambayana S., S.Pd., M.Pd. .2017. An analysis of jargons used by receptionsis in front office at GRAND ISTANA RAMA Hotel. Jurnal Pendidikan Bahasa Inggris undiksha.Vol.5, No.2, 2017. http://dx.doi.org/10.23887/jpbi.v5i2.13611

[2] Ratna Sari Dewi. 2019. The analysis of jargon used by WhatsApp community among Tarbiyah Lecturers in UINSU. The 1st Multi-Disciplinary International Conference University of Asahan2019. North Sumatra, Indonesia, March 23th ,2019. 1151-1158.

[3] Satoshi Abiko, Dai Hasegawa, Michal Ptaszynski, Kenji Nakamura, and Hiroshi Sakura. 2018. Method for estimation of harmfulness of ID-Exchange BBS based on lexical jargonizations. Journal of the Information Systems Society of Japan (JISSJ). Vol.13, No.2. 41-58. https://doi.org/10.19014/jissj.13.2_41

[4] Hiroshi Ohnishi, and Keishi Tajima. 2013. Discovering new jargons based on skew of word appearance distribution. DBSJ Journal.Vol.12, No.1. 49-54.

[5] Tatsuya Aoki, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2017. Distinguishing Japanese nonstandard usages from standard ones. In Proceedings of EMNLP'17. Copenhagen, Denmark, September 7-12th, 2017. 2323-2328.

[6] Tatsuya Aoki, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2019. Detecting nonstandard word usages on social media. NLP Journal. Vol. 26, No.2. 381-406. https://doi.org/10.5715/jnlp.26.381

[7] Jose Marcio Duarte, Samuel Sousa, Evangelos Milios, and Lilian Berton. 2021. Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations. Information Sciences. Vol.570. 278-297. https://doi.org/10.1016/j.ins.2021.04.006

[8] Seyed Mahdi Rezaeinia, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi. 2019. Sentiment analysis based on improved pre-trained word embeddings. Expert Systems with Applications.Vol.117. 139-147. https://doi.org/10.1016/j.eswa.2018.08.044

[9] Kangzhi Zhao, Yong Zhang, Chunxiao Xing, Weifeng Li, and Hsinchun Chen. 2016. Chinese underground market jargon analysis based on unsupervised learning. 2016 IEEE Conference on Intelligence and Security Informatics: Cybersecurity and Big Data (ISI). Tuson, USA, September 28-30, 2017. https://doi.org/10.1109/ISI.2016.7745450

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality, In Proceedings of Neural Information Processing Systems 26. Harrahs and Harveys, Lake Tahoe, USA, December 5-10,2013. 2123-2131.

[11] Ryo Inui, and Satoshi Yamamura. 2019. Processing of split writing characters using visual "reading". In Proceedings of the Association for Natural Language Processing. Nagoya, Japan, March 12-15, 2019. 434-437.