International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

Contextual Word2Vec Model for Understanding Chinese Out of Vocabularies on Online Social Media

JiaKai Gu, Chung-Ang University, South Korea
https://orcid.org/0000-0001-9913-7864

Gen Li, Chung-Ang University, South Korea

Nam D. Vo, FPT University, Vietnam
https://orcid.org/0000-0001-5469-5707

Jason J. Jung, Chung-Ang University, South Korea*

ABSTRACT

In this chapter, the authors propose to use contextual Word2Vec model for understanding OOV (out of vocabulary). The OOV is extracted by using left-right entropy and point information entropy. They choose to use Word2Vec to construct the word vector space and CBOW (continuous bag of words) to obtain the contextual information of the words. If there is a word that has similar contextual information to the OOV, the word can be used to understand the OOV. They chose the Weibo corpus as the dataset for the experiments. The results show that the proposed model achieves 97.10% accuracy, which is better than Skip-Gram by 8.53%.

KEYWORDS

Out of Vocabulary (OOV), Social Media, Word Embedding, Word2Vec

INTRODUCTION

The understanding of textual documents is based on the semantic meaning of each word in the natural language when studied via social networks, machine translation, information extraction, sentiment analysis, text classification, and other semantic-based natural language processing research. However, the semantics are generally unclear to a computer due to the existence of a high number of out-of-vocabulary (OOV) terms. Therefore, a semantic understanding of OOV is an obstacle to overcome in the field of natural language processing.

The China Internet Network Information Center's (CNNIC, 2022) 49th Statistical Report on the Development Status of China's Internet showed that the size of China's Internet users reached 1.032

DOI: 10.4018/IJSWIS.309428 *Corresponding Author

billion as of December 2021. This number was an increase of 42.96 million users from December 2020. Such a large user base provides a rich corpus for Chinese natural language processing-related research.

A Chinese sentence includes several consecutive words. To understand the semantics of Chinese, it is necessary to divide the sentence into strings of words, with each word serving as a basic unit. There is no obvious separation between words in Chinese; therefore, the wrong separation can lead to ambiguity (Blythe et al., 2012). This creates challenges in finding OOV in text (see Figure 1). As a phenomenal short text-based real-time social network, Twitter provides a rich corpus of information for natural language processing (Murthy et al., 2019). Similarly, to study Chinese OOV, Ahmed et al. (2022) noted that social media platforms provide unique opportunities for conducting social science and Web-based research. Therefore, this study chose the Twitter-like Chinese media social network, Weibo (https://www.weibo.com), as its corpus (Zhu et al., 2021). Weibo contains many Chinese colloquialisms and slang, which provides useful information when exploring OOV and semantics.

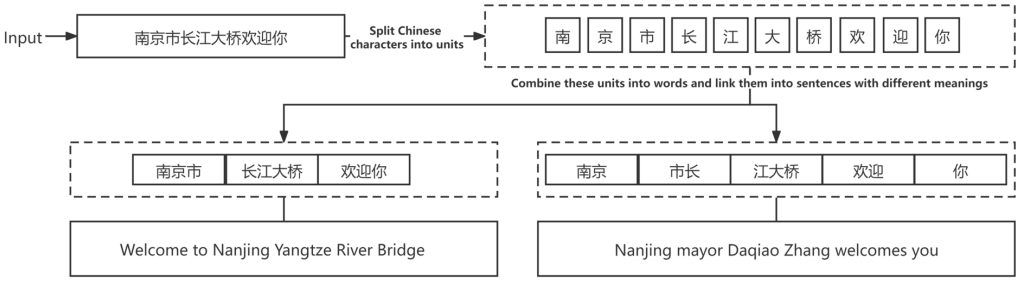
In this work, the authors seek to extract Chinese OOVs from social networks to understand the meaning of Chinese OOVs in social networks. The authors propose to use information from the context to understand OOVs. The current study was inspired by Nagy et al. (1987), in which context was used to understand the meaning of words. To utilize this approach to understanding word meaning, the authors must extract, analyze, and use relevant contextual information from the OOV. The word2vec's continuous bag of words (CBOW) model can capture valid contextual information to calculate word similarity and understand word meanings.

Most OOV-related research uses a large corpus and named entity extractions. However, there are fewer studies on OOV of social networks and their lexical meanings. The contributions of this research include:

- Limited amount of relevant social network content as a corpus, which addresses the problem of sparse low-resource linguistic corpora
- Contextual information of OOV to quickly understand the meaning of OOV words, which facilitates semantic and natural language processing-related research.
- Use the content of people via OOV in social networks as a corpus, which makes its semantic features more controllable and effective for OOV meaning understanding

The next section presents research related to Chinese OOV detection and the meaning of OOV understanding. To better understand the study's proposed method, the research introduces the underlying theories of information entropy and distribution representation. Then, the article provides the methodology of the proposed approach for extracting and understanding Chinese OOV through social networks. This is followed by an illustration of the data through relevant experiments. It shows

Figure 1. Examples of ambiguity in the same Chinese sentence with different word separation



the results and discusses the impact of two word2vec models on the current study's experimental results. Finally, the article presents the conclusions, limitations of the current research, and prospects for future research.

BACKGROUND

Chinese OOV Detection

Word segmentation is usually regarded as an integral aspect of the Chinese natural language process approach to OOV detection. Peng et al. (2004) noted that conditional random fields have been presented as a typical paradigm. Chinese word segmentation and new word detection are combined in a hybrid model presented by Sun et al. (2012). New high-dimensional features, such as word-based features and enhanced edge (label-transition) features, are presented for the combined modeling of words and transitions. Supervised models require relevant annotations to operate. Therefore, in some cases without annotations, it is impossible to detect OOV accurately. There are also unsupervised models. Sun et al. (2012) proposed top-down word discovery and segmentation, an unsupervised tool for discovering new Chinese words and sentences, rating words, and segmenting text.

Combining systems with contextual analysis tools to create a pipeline enables researchers to gain insight into OOV without training. While this has been shown to be an effective type of experiment, this class of models requires significant time to compute input documents and process results. There are also challenges in user behavior. Zheng et al. (2009) discussed issues related to collecting users, typing habits, and probability calculations in extracting OOV. However, the method is feasible because this type of commercial data is difficult to obtain.

UNDERSTANDING THE MEANING OF CHINESE OOV

In previous works, Glyph2Vec (Chen et al., 2020) was used to study the semantics of OOV, as well as analyze the meaning of words through the shape of Chinese characters (e.g., 飯 (meal) → 食 (eat)+ , 話 (talk) → 言 (speech)+舌 (tongue)). This method does not require significant corpus support. However, the method is only applicable to traditional Chinese and pictographic characters. OOVs broadly used in recent social networks are characterized by diversity and innovation. It uses simplified words to express the meaning of words. In addition, it includes some new meanings of old words.

Jianju and Feng (2018) utilized a knowledge base to comprehend OOV and the meaning comprehension of words through morphemes inside words and word structure in an established knowledge base. This approach combines a knowledge base with a corpus to train a model of overlapping words. It uses a staged algorithm to automatically understand the morphemic constructions of OOV knowledge. However, it also requires the support of its large understanding and knowledge base.

The many types of word-formation knowledge and unstable parts of speech in social networks impact the results of this method. The semantic understanding of OOV based on word-formation knowledge is well-founded. However, due to the existence of polysemy and irregular word-formation in the Chinese language, it is difficult to understand OOVs by word-formation knowledge without additional information. Chen and Chen (2000) proposed affixes to distinguish meanings of words. However, modern OOV features do not allow the meaning of words to be understood by affixes alone.

Another method uses part-of-speech (POS) to annotate words with lexical properties and other content (Qu et al., 2021). However, there are few POS tags in dynamically formed OOV communication environments like social networks. Additionally, there are limited corpora and repositories for support for some low-resource languages.

INFORMATION ENTROPY

Entropy is a measure of the quantity of information in a system (Tsai et al., 2008). Higher entropy indicates greater information richness, increased uncertainty, and more difficulty in guessing. Typically, the entropy of a random variable *S* could be stated as:

$$H\left(S\right) = -\sum_{s\in S} p\left(s\right)\log_2 p\left(s\right) \tag{1}$$

p(*s*) denotes the probability of occurrence of the event *s*, which is the probability of occurrence of a word in the case of OOV mining. Therefore, information entropy may be used to quantify the quantity of data. The information entropy of a candidate is evaluated through the word's left and right sides to determine if it has a strong left-right collocation. In addition, it studies whether a particular threshold is met. This study will consider two fragments to be an OOV.

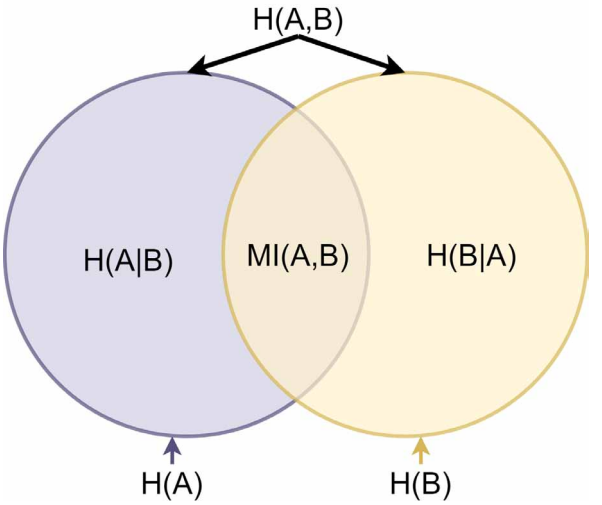
Mutual information (MI) is a useful measure of information in information theory (Hao et al., 2022; Steuer et al., 2002). It may be thought of as the amount of information in a random variable about another random variable. It may also be the decrease of uncertainty in a random variable because of knowledge about another random variable.

The entropy represents the amount of information. As shown in Figure 2, MI represents the amount of information shared by random variables. Given variables *A* and *B*, it is formulated as:

$$MI\left(A,B\right) = \sum_{a\in A, b\in B} p\left(a,b\right)\log_2 \frac{p\left(a,b\right)}{p\left(a\right)p\left(b\right)}. \tag{2}$$

It can be said that MI represents the reduction in uncertainty about either variable after knowing the other. It can also be said that MI represents the reduction in uncertainty about the other variable after knowing either variable.

Figure 2. Relationship between MI and information entropy



DISTRIBUTIONAL REPRESENTATION HYPOTHESIS

The distributional representation hypothesis is based on words with similar contexts having similar semantics. Bengio et al. (2003) proposed a neural network-based language model (NNLM) and introduced the concept of "word embedding." Afterward, word vectors were proposed by He et al. (2018) to train the surrounding context of a target word and carry contextual information about the word (Jiang & He, 2020; Lan et al., 2021).

Barnickel et al. (2009) proposed the SENNA model for the purpose of generating word vector representations. However, all these models suffer from long training times. Mikolov et al. (2013) proposed a CBOW and skip-gram model that can be trained efficiently under a large scale by simplifying the network and designing accelerated training methods based on the previous neural network models. Word vectors generated by CBOW and skip-gram can capture the semantic correlation between words. The law of linear operations between word vectors can also be found. The better performance of CBOW and skip-gram on multiple tasks, as well as the efficiency of training large corpora, make them milestone models in neural network word representation learning.

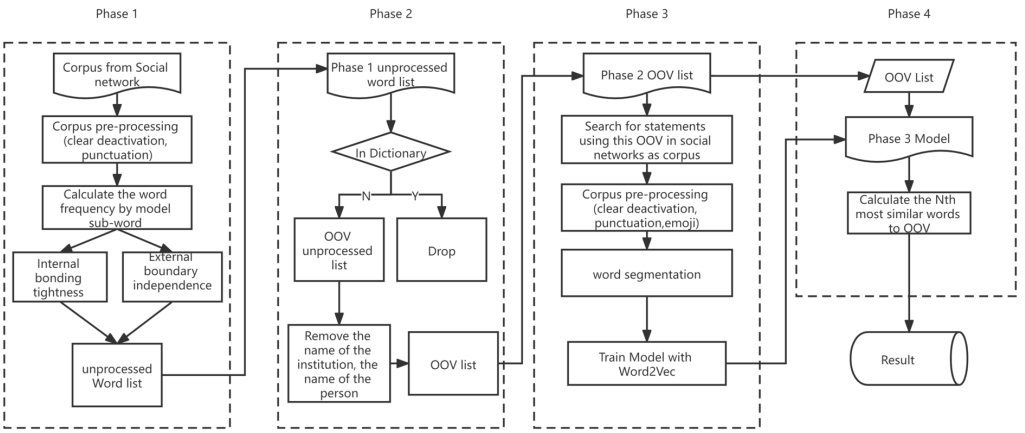
This study discusses the use of a small or relatively real-time corpus to obtain recent OOVs from the social network of microblogs. It uses the OOVs to mine text containing the OOVs from the social network. The meaning of words can be understood through context. The word2vec can train vector matrices that contain a large amount of contextual information for use. Therefore, the authors use the word2vec technique to train word vectors and understand the meaning of an OOV by exploiting the contextual information of that OOV.

METHODOLOGY

The comprehension procedure of Chinese OOV is divided into four stages, as shown in Figure 3.

The initial stage uses a corpus of microblogs. This is processed to extract words. The second stage compares the extracted words from the first stage using an Internet dictionary to obtain a list of OOVs. In the third stage, OOVs are filtered by the second stage comparison. Keywords are used to extract text data containing the OOVs in Weibo as a corpus. This is merged with the original corpus for preprocessing. Then, it trains the word vector model. The final stage calculates the *N* words that mirror the OOV. It also outputs the results.

Figure 3. Program work



Possible words in the corpus must be extracted to alternatively filter the OOV. A string must satisfy the following four properties to be called a word: (1) circulation; (2) completeness; (3) flexibility; and (4) cohesion (Kam-Fai et al., 2009). Therefore, before understanding the meaning of OOV, words must be extracted from a sentence in the corpus. Words in a Chinese sentence are not separated by spaces; therefore, it is difficult to extract OOV from Chinese sentences using traditional word separation tools like THULAC (Li & Sun, 2009) or LTP (Che, 2020).

There are several steps to extract a word from a corpus. The first step selects candidate words via word frequency. N-Gram is used to segment the corpus to obtain word fragments (Cavnar & Trenkle, 1994). The word fragment's frequency is counted and a threshold is set. The word fragment constitutes a candidate word if only the frequency exceeds it. The second step distinguishes between the internal cohesion and external boundary independence of the candidate words. It selects statistics based on the degree of cohesion within the word and boundary measurement outside the word. Then, a specific screening method calculating point MI (PMI) and left-right entropy are applied (Wang et al., 2020).

As discussed, MI is the calculation of a mean value for a random variable. The computer field tends to use PMI because it calculates the MI between two specific events. PMI is used to determine the degree of tightness of union within words. This method is based on the probability function of two events occurring at the same time. It reflects the degree of interdependence between two variables. Let the joint distribution of two random variables (x, y) be $P(x, y)$, the edge distribution is $P(x)$, $P(y)$, their $PMI(x, y)$ is the relative entropy of joint distribution $P(x, y)$, and the edge distribution is $P(x)$, $P(y)$. The higher the value of PMI, the higher the possibility of a boundary between X and Y, as shown in equation 3.

$$PMI\left(x;y\right) = log\frac{p\left(x,y\right)}{p\left(x\right)p\left(y\right)} = log\frac{p\left(x \mid y\right)}{p\left(x\right)} = log\frac{p\left(y \mid x\right)}{p\left(y\right)} \tag{3}$$

To ensure that OOV words have legal semantics, it is necessary to ensure that OOV words are independent language units. Information entropy can indicate the quantity of information. The study evaluates the information entropy of a candidate word's left and right sides to determine if it has rich left-right collocation. If a particular threshold is met, the study considers two fragments to be a new word. Left-right entropy refers to the entropy of vocabulary's left boundary $E_L(W)$ and right boundary $E_R(W)$ defined by:

$$E_L\left(W\right) = -\sum\nolimits_{\forall a \in A} P\left(aW \mid W\right) * log_2 P\left(aW \mid W\right) \tag{4}$$

$$E_R\left(W\right) = -\sum\nolimits_{\forall b \in B} P\left(Wb \mid W\right) * log_2 P\left(Wb \mid W\right) \tag{5}$$

where W indicates candidate words after the N-Gram segmentation is formulated as $W = \left\{w_i | i \in \left(0, n\right)\right\}$. A is a collection of all words appearing on the left of a candidate, a is a word appearing on the left, B is a collection of all words appearing on the right of a candidate, and b is a word appearing on the right. If the values of E_L and E_R for a candidate word are larger, the more words appear around the candidate word W. Therefore, the more likely it is that W is a word. Both MI and left-right entropy factors are indispensable. If only MI is used, only half of the words will be recognized. If only left-right entropy is used, the value of some articulated words will be very high.

The candidate list constructed by the above method is matched through the Baidu Chinese dictionary. If the word is not found in the dictionary, it is added to the OOV list. Although most of the names of entities are not registered in the dictionary, it is assumed that named entities do not belong to new words. Therefore, named entities are filtered out by their features. A cleaned OOV list is obtained. In addition to using the symbol library to remove punctuation, most of the emoticons are removed using regular expressions (Burhanuddin & Muhammad, 2019). In the extraction stage of OOV, Chinese stop words are used. The jieba tool is used in the semantic understanding phase (https://github.com/fxsjy/jieba). The extracted OOV is added to a user-defined dictionary to enhance ambiguity correction and achieve accurate Chinese word segmentation.

The aim is to explain the lexical meaning of OOV through context. The sliding window word characterization method determines the number of contextual words around the target word. After establishing the contextual information of the target words, the probability of the target word is maximized using the objective function of the language model. As Figure 4 shows, each word in the sentence becomes a target word once the window is moved. When the window is moved to the next, the target word in the current window becomes the context information for the target word in the next window. Therefore, the same word can serve as both the target word and the context information during the algorithm learning process. Given that the same word might have two distinct states, the algorithm simultaneously learns two-word vector matrices, one to represent the word vector while it is the target word and another to represent the word vector when it is context information. Using rich contextual information is helpful in understanding the meaning of OOV.

In wod2vec, CBOW defines the words within a certain window to the left and right of the central word as the context. The idea is to use the modeled context to predict the central word by the log-linear classification model. NNLM is found to be costly in training time due to its nonlinear hidden layer. CBOW discards the nonlinear hidden layer. In addition, unlike the NNLM mapping layer, CBOW no longer uses word vector stitching. CBOW does not consider the order of words in the context; therefore, it is called the continuous bag-of-words model. The continuous refers to the distributed contextual representation of the continuous space.

Given an input sequence $W = \left\{w_i | i \in \left(0, n\right)\right\}$, CBOW maps the context c through the mapping layer e. The words in the context are mapped to the representation of the context, defined by:

Figure 4. Example of sliding window

$$h = \frac{1}{n-1}\sum_{w_i \in c} e\left(w_i\right) \tag{6}$$

The contextual representation *hh* is then used to predict the central word by maximizing the conditional probability. It is defined by:

$$L = \sum_{(w,c)\in D} logP(w \mid c) \tag{7}$$

$$P(w \mid c) = \frac{exp(e'\left(w\right)^T h\right)}{\sum_{w' \in v} exp(e'\left(w'\right)^T h\right)} \tag{8}$$

where w is the target word, w_j is each word in context information c.

According to the properties of CBOW, the calculation of contextual words to central words is obtained by the trained matrix after getting the word vector of each word. It uses known vocabulary from similar contexts to understand OOV and calculate the k words with the largest similarity in the corpus as its prediction candidate set. In this work, when two words have similar or identical contexts, it is sufficient to use one word to understand the other word. This study draws inspiration from Ismail et al.'s (2022) use of cosine similarity, calculating the distance between the two-word vectors OOV O and Word W to find their relationship where n is the dimensionality of the word vector. Figure 5 shows an example for understanding the relationship between the O and W.

$$\mathrm{sim}\left(O,W\right) = \frac{\sum_{i=1}^{n} O_i W_i}{\sum_{i=1}^{n} O_i^2 \sum_{i=1}^{n} W_i^2} \tag{9}$$

For the example OOV "新冠 (COVID-19)," n = 20, Table 1 shows the 20 most similar words.

Figure 5. Example of the relationship of O and W

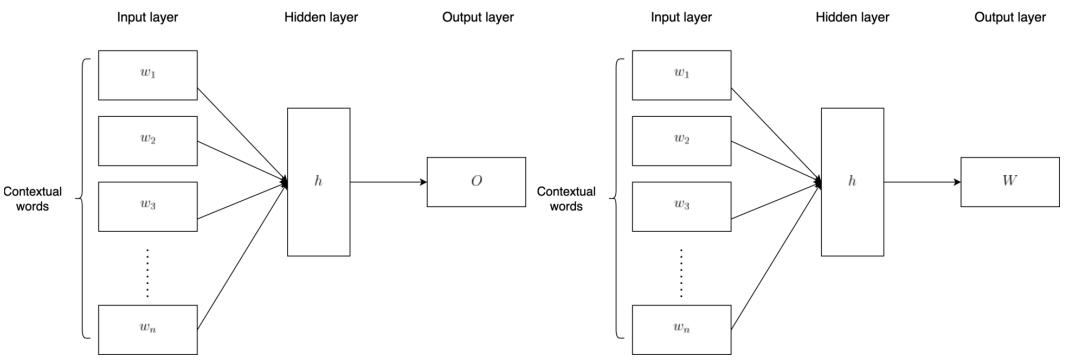


Table 1. 20 Most similar words of "新冠 (COVID-19)"

Word	Translation	Word	Translation
感染	Infection	爆发	Outbreak
病毒	Virus	钢琴家	Pianist
肺炎	Pneumonia	流感	Influenza
世卫	WHO	傅聪	Cong Fu (Name)
HPV	HPV	注射	Injection
变异	Mutation	呈	Present
韦尔	Weil	挪威	Norway
疫苗	Vaccine	钟南山	Ns. Zhong (Name)
一家医院	A hospital	未感染	Uninfected
变种	Variant	血凝	Hemagglutination

EXPERIMENT AND DISCUSSION

Experimental Environment

In the experiment, the system was built with the Python 3.9.1 programming language. Data collection used the request package of Python, word2vec model training by Gensim 4.1.2 of Python, and parameter setting of Gensim. The study used the default preset values for the experiments. The system was implemented on a computer with an Apple M1 processor, 4×3.2Ghz + 4×2.064Ghz CPU, 8GB RAM, and Mac OS big sur 11.4 operating system. The system source code was uploaded in GitHub (https://github.com/gabrielpondc/oovunderstand); some detail was uploaded through the project Website (http://recsys.cau.ac.kr:8095).

DATASET

This study used 10,308 Weibo articles published by a news author account from November 11, 2020, to May 4, 2021 (Corpus A). There were 1,887,142 Chinese characters in Corpus A (https://www.weibo.com/breakingnews). In addition, a total of 3,184 words were isolated from the corpus by MI and left-right entropy. A total of 311 OOVs were extracted from the words. Thirty-five OOVs existed after filtering the named entity with characteristics.

In general, word vectors can be trained through a large corpus. However, this method does not work for low-frequency words. In addition, OOV tends to be low-frequency words, which are trained with sparse data and do not get better word vectors. The study used a total of 4,322 articles with the OOVs extracted from Weibo via the excavated OOVs (Corpus B). After processing, there were 634,136 Chinese characters in Corpus B. The word2vec models were trained using Corpus A and Corpus B.

EXPERIMENTAL RESULTS

The test set of 35 OOVs were understood. For understanding results, the following three categories were evaluated with a manual annotation. The first category, (**A**), is the word with the highest similarity to the OOV that can be directly and semantically substituted like the OOV 天才病 (genius disease) for which the word with the highest similarity is 阿兹伯格综合症 (Asperger's syndrome). The second category, (**B**), is the number of words with high similarity to the OOV that are associated with the OOV. For instance, the OOV 新冠 (COVID-19) has three words with high similarity: 感染

(infection), 病毒 (virus), and 肺炎 (pneumonia). The third category, (**C**), is the OOV that is unrelated to the words whose similarity results are larger. For example, the OOV 凤凰网 (media organization) has three words with larger similarity: 应该 (should be), 讨论 (discuss), and 看法 (view).

Table 2 lists the accuracy calculated by equation 10, where *n* is the number of OOVs. The accuracy of skip-gram is better than CBOW in most of the studies. However, due to the small corpus in this experiment, the accuracy of CBOW is higher than skip-gram. In Category B, the use of skip-gram is more accurate than the interpretation of CBOW. For instance, 居家 refers to things associated with living or staying at home. If using CBOW as a model, the three most similar words are 划算 (value for money), 日常 (daily), and 租房 (rent a home). If using skip-gram, the three most similar words are 生活 (living), 隔离 (isolation), and 办公 (office).

$$Accuracy = \frac{\left(A + B\right)}{n} \times 100\% \tag{10}$$

DISCUSSION

Like the CBOW model, the skip-gram model discards the time-consuming nonlinear hidden layer, making the model simpler and able to adapt to large-scale training data. However, in contrast to the CBOW model, skip-gram chooses to use central words to predict words in context.

To prove the validity of the experiment, the additional experiment decided to discard the original corpus and construct a temporary corpus to understand word meanings by searching on Weibo through OOV. The temporary corpus used in the retroactive experiment was de-duplicated with 331 posts and 18,000 Chinese characters by 耗子尾汁 (a harmonic of 好自为之). The actual meaning of OOV was finally obtained by removing the adjuncts and named entities like 好自为之 (You are on your own, most of the time when you're disappointed in someone).

CBOW uses context to understand words. Skip-gram uses words to understand context. Therefore, CBOW is more suitable for understanding OOV when the current corpus is not large. It can also be seen in the results that the similarity between 耗子尾汁 and 好自为之 using the CBOW model is 0.1. The similarity between 耗子尾汁 and 好自为之 of the skip-gram model is 0.92. The example is shown in Figure 6.

The skip-gram operational model uses central words to make predictions about the context. It does not help to understand the meaning of OOV. The results of CBOW are more suitable for understanding OOV. Regarding skip-gram, a larger corpus is more helpful for the experimental results; however, it is difficult to have a large corpus to support the experiments due to the usage characteristics of OOV.

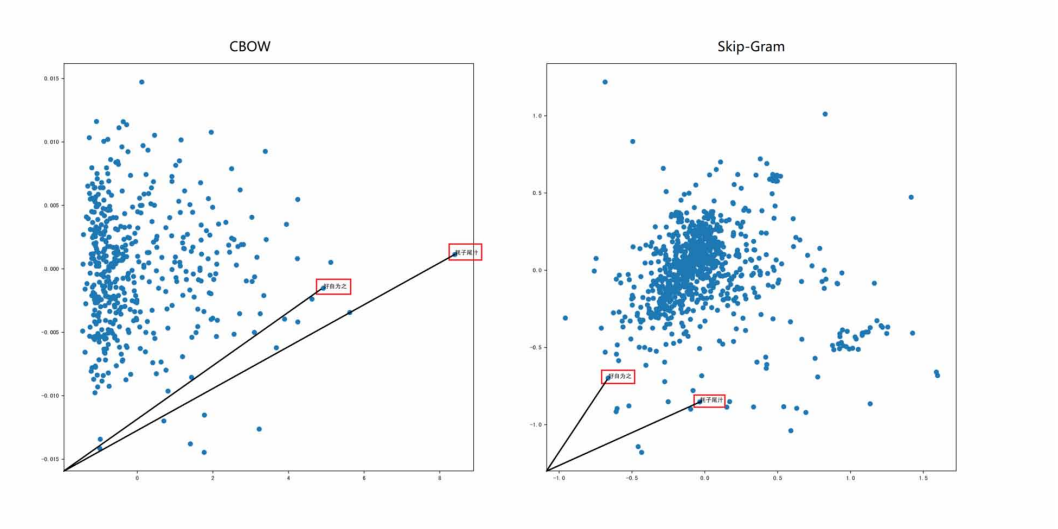
CONCLUSION

The Chinese OOV comprehension program uses four phases to effectively solve the problem of semantic understanding of Chinese OOV by machines in natural language processing. These include

Table 2. Caption should be sentence case with no ending punctuation if only one sentence

Model	A	B	C	Accuracy
CBOW	21	13	1	97.10%
Skip-Gram	17	14	4	88.57%

Figure 6. Understanding of 耗子尾汁 by CBOW and skip-gram models



mining data as corpus, recognizing OOV, building a corpus by using OOV, and understanding OOV. It has also played an important role in expanding the Chinese lexicon.

Understanding the semantic meaning of OOV is important in natural language processing research. This article uses MI and left-right entropy to discover OOVs by mining the texts from social networks in combination with online dictionaries. It applies contextual word vectors to reflect external features of words to the semantic understanding of OOVs.

The experimental result's accuracy rate can reach 97.10%. The accuracy of the CBOW trained model was 8.53% higher than the skip-gram trained model in the case of a small corpus. However, the results were not satisfactory for some words with high contextual noise. The understanding of specialized words and new use of ancient words achieves underwhelming results. Such OOV words must be supplemented by the combination of internal word information.

There are concerns with the findings as time changes. Dictionary contributors regularly upload new words, causing the OOV in this library to cease to be OOV and become a known word. Still, the article's focus on the collection and understanding of OOV is not affected.

The skip-gram model is more accurate than the CBOW model for the OOV in category B because the category B word appears less often in the corpus. This makes the contextual information redundant and cumbersome. The feature of skip-gram performs multiple operations on each input OOV center word in the word, which optimizes the output result. The next step is to introduce more external information, such as a knowledge graph, to improve the understanding of the semantic meaning of OOVs.

CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (NRF-2020R1A2B5B01002207, NRF-2021R1I1A1A01060302).

REFERENCES

Ahmed, S., Rajput, A., Sarirete, A., & Chowdhry, T. J. (2022). Flesch-Kincaid measure as proxy of socio-economic status on Twitter: Comparing US senator writing to Internet users. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–19. doi:10.4018/IJSWIS.299858

Barnickel, T., Weston, J., Collobert, R., Mewes, H.-W., & Stümpflen, V. (2009). Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS One*, *4*(7), e6393. doi:10.1371/journal.pone.0006393 PMID:19636432

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., & Gauvain, J.-L. (2003). Neural probabilistic language models. In D. E. Holmes & L. C. Jain (Eds.), *Innovations in machine learning: Theory and applications* (pp. 137–186). Springer. doi:10.1007/3-540-33486-6_6

Blythe, H. I., Liang, F., Zang, C., Wang, J., Yan, G., Bai, X., & Liversedge, S. P. (2012). Inserting spaces into Chinese text helps readers to learn new words: An eye movement study. *Journal of Memory and Language*, *67*(2), 241–254. doi:10.1016/j.jml.2012.05.004

Burhanuddin, A., & Muhammad, H. (2019). The language of emoji in social media. *KnE Social Sciences*, *3*(19). Advance online publication. doi:10.18502/kss.v3i19.4880

Cavnar, W. B., & Trenkle, J. M. (1994). *N-gram-based text categorization*. https://www.semanticscholar.org/paper/N-gram-based-text-categorization-Cavnar-Trenkle/49af572ef8f7ea89db06d5e7b66e9369c22d7607

Che, W., Yunlong, F., Qin, L., & Liu, T. (2020). *A open-source neural Chinese language technology platform with pretrained models.* arXiv preprint arXiv:2009.11616.

Chen, H.-Y., Yu, S.-H., & Lin, S.-d. (2020). Glyph2Vec: Learning Chinese out-of-vocabulary word embedding from glyphs. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. doi:10.18653/v1/2020.acl-main.256

Chen, K.-J., & Chen, C.-j. (2000). Automatic semantic classification for Chinese unknown compound nouns *Proceedings of the 18th conference on Computational linguistics* (vol. 1). doi:10.3115/990820.990846

China Internet Network Information Center (CNNIC). (2022). *49th statistical report on the development of the Internet in China*. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202202/P020220318335949959545.pdf

Hao, Y., Wang, L., Liu, Y., & Fan, J. (2022). Information entropy augmented high density crowd counting network. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–15. doi:10.4018/IJSWIS.297144

He, L., Du, Y., & Zhang, L. (2018). *Vector representation of words for detecting topic trends over short texts*. Academic Press.

Ismail, S., Shishtawy, T. E. L., & Alsammak, A. K. (2022). A new alignment word-space approach for measuring semantic similarity for Arabic text. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–18. doi:10.4018/IJSWIS.297036

Jiang, D., & He, J. (2020). Tree framework with BERT word embedding for the recognition of Chinese implicit discourse relations. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 162004–162011. doi:10.1109/ACCESS.2020.3019500

Jianju, Q. U., & Feng, M. (2018). Sense prediction of Chinese unknown words based on knowledge base. *Journal of Chinese Information Processing*.

Kam-Fai, W., Wenjie, L., Ruifeng, X., & Zheng-sheng, Z. (2009). *Introduction to Chinese natural language processing*. Morgan & Claypool. https://ieeexplore.ieee.org/document/6813529

Lan, Y., He, S., Liu, K., Zeng, X., Liu, S., & Zhao, J. (2021). Path-based knowledge reasoning with textual semantic information for medical knowledge graph completion. *BMC Medical Informatics and Decision Making*, *21*(9), 335. doi:10.1186/s12911-021-01622-7 PMID:34844576

Li, Z., & Sun, M. (2009). Punctuation as implicit annotations for Chinese word segmentation. *Computational Linguistics*, *35*(4), 505–512. doi:10.1162/coli.2009.35.4.35403

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781.

Murthy, J. S., G. M., S., & K. G., S. (2019). A real-time Twitter trend analysis and visualization framework. *International Journal on Semantic Web and Information Systems*, *15*(2), 1–21. doi:10.4018/IJSWIS.2019040101

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, *24*(2), 237–270. doi:10.3102/00028312024002237

Peng, F., Feng, F., & McCallum, A. (2004). Chinese segmentation and new word detection using conditional random fields *Proceedings of the 20th International Conference on Computational Linguistics*. doi:10.3115/1220355.1220436

Qu, S., Liu, W., Li, J., & Peng, Z. (2021). An enhancement method for Chinese environment semantic slot filling based on POS tagging. *2021 International Conference on Control, Automation and Information Sciences (ICCAIS)*. doi:10.1109/ICCAIS52680.2021.9624600

Steuer, R., Kurths, J., Daub, C. O., Weise, J., & Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics (Oxford, England)*, *18*(suppl_2), S231–S240. doi:10.1093/bioinformatics/18.suppl_2.S231 PMID:12386007

Sun, X., Wang, H., & Li, W. (2012). Fast online training with frequency-adaptive learning rates for Chinese word segmentation and new word detection. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers* (vol. 1).

Tsai, D.-Y., Lee, Y., & Matsuyama, E. (2008). Information entropy measure for evaluation of image quality. *Journal of Digital Imaging*, *21*(3), 338–347. doi:10.1007/s10278-007-9044-5 PMID:17577596

Wang, G., Tao, Y., Ma, H., Bao, T., & Yang, J. (2020). Research on key technologies of knowledge graph construction based on natural language processing. *Journal of Physics: Conference Series*, *1601*(3), 032057. doi:10.1088/1742-6596/1601/3/032057

Zheng, Y., Liu, Z., Sun, M., Ru, L., & Zhang, Y. (2009). Incorporating user behaviors in new word detection. *Proceedings of the 21st International Joint Conference on Artificial Intelligence*.

Zhu, Y., Li, X., & Wang, J. (2021). Analysis and research of Weibo public opinion based on text. *Journal of Physics: Conference Series*, *1769*(1), 012018. doi:10.1088/1742-6596/1769/1/012018

Jiakai Gu received the B.Eng in the School of Computer Science and Engineering from Chung-Ang University, Seoul city, Korea, in 2021. He is the Master candidate in the School of Computer Science and Engineering from Chung-Ang University, Seoul city, Korea from 2021. His research interests are data mining, deep learning, word embedding and nature language processing. Recently, he has been working on understanding and translation the meaning of Out-Of-Vocabularies form Social Networks.

Gen Li received the B.Eng in the mechanical Design manufacture and Automation from Haidu College Qingdao Agricultural University, Yantai city, China, in 2015. He received his MS degree in mechanical engineering from Kunsan University, Kunsan city, South Korea, in 2017. He received his Ph.D. from the department of computer engineering, Chung-Ang University, Seoul, South Korea, in 2022. His research interests are data mining, deep learning, anomaly detection, graph embedding, and time- series analytics. For example, his recent research is related to anomaly detection from the multivariate time series based on graph embedding.

Nam D. Vo is a postdoctoral researcher at Knowledge Engineering and Storytelling Laboratory, Chung-Ang University since September 2020. He received a B.S. degree in Information Technology from Danang University, Vietnam, in 2005; an M.S. degree from Nice Sophia Antipolis University, France in 2011, and a Ph.D. degree in Application Software from Chung-Ang University, South Korea, in August 2020. His research interests include cross-domain recommendation systems and user pattern detection.

Jason J. Jung is a Full Professor in Chung-Ang University, Korea, since September 2014. Before joining CAU, he was an Assistant Professor in Yeungnam University, Korea since 2007. Also, he was a postdoctoral researcher in INRIA Rhone-Alpes, France in 2006, and a visiting scientist in Fraunhofer Institute (FIRST) in Berlin, Germany in 2004. He received the B.Eng. in Computer Science and Mechanical Engineering from Inha University in 1999. He received M.S. and Ph.D. degrees in Computer and Information Engineering from Inha University in 2002 and 2005, respectively. His research topics are knowledge engineering on social networks by using many types of AI methodologies, e.g., data mining, machine learning, and logical reasoning. Recently, he has been working on intelligent schemes to understand various social dynamics in large scale social media.