An OOV Recognition Based Approach to Detecting Sensitive Information in Dialogue Texts of Electric Power Customer Services

Xiao Liang[1](✉), Ningyu An[1], Ning Wu[2], Yunfeng Zou[2], and Lijiao Zhao[1]

[1] Global Energy Interconnection Research Institute Co. Ltd.,
Artificial Intelligence on Electric Power System State Grid Corporation
Joint Laboratory (GEIRI), Beijing 102209, China
33180900@qq.com
[2] State Grid Jiangsu Electric Power Co., Ltd. Marketing Service Center,
Nanjing 210019, China

Abstract. Sensitive word recognition technology is of great significance to the protection of enterprise privacy data. In electric power custom services systems, the dialogue texts recording the conversational information between electric power customers and the customer services staffs contain some sensitive information of electric power customers. However, the colloquialism and synonyms in dialogue texts often make sensitive information recognition more difficult. In this paper, we proposed an out-of-vocabulary (OOV) approach for recognizing sensitive words in the dialogue texts of electric power customer services. We combine the semantic similarity based on word embeddings and structural semantic similarity based on HowNet for recognizing sensitive OOV words in the dialogue texts. The related experiments were made, and the experimental results show that our method has higher recognition accuracy in comparison with the popular approaches.

Keywords: Out-of-vocabulary · Sensitive word recognition · HowNet · Word embedding · Electric power customer services

1 Introduction

Sensitive words generally refer to those words that possibly are unhealthy and uncivilized words with political tendency and violent tendency. Sensitive word recognition refers to the technology of detecting and recognizing these sensitive words from the original documents so that they can be further processed. Sensitive word recognition technology is of great significance to the protection of enterprise privacy data. For example, in the electric power system, the collected data containing customer personal information will be used for analysis and research. If these data are used publicly, it will reveal customer privacy and even bring danger to customers. In the situation, it is necessary to shield customer personal information to prevent customer information from being leaked. Therefore, it is very important to identify and process sensitive words in data. In customer services, sensitive words often are some personal

S. Shi et al. (Eds.): AICON 2020, LNICST 356, pp. 598–607, 2021.
https://doi.org/10.1007/978-3-030-69066-3_52

privacy information during the conversation between customer service staffs and clients, such as name, affiliations, personal ID, card numbers and other private information. Just because of their privacy, these sensitive words should be recognized and further processed before the related data is available to the public by eliminating or masking some private information.

In the past decades, some methods for processing sensitive words have been proposed based on some natural language processing. They have been widely used in the field of sensitive word recognition successfully. Zhou and Gao [1] applied the double Hash method for open addressing to string matching, and built a secondary Hash table, which kept the hash value calculation efficient and improved the recognition accuracy, but this matching method can only recognize specific sensitive words. Yu et al. [2] proposed a decision tree based sensitive word recognition algorithm, which established a decision tree based on a sensitive vocabulary, and improved the accuracy of sensitive words detection through the multi-factor perspectives. However, these methods are lack of semantic analysis between words, and have to manually judge the forward and reverse meanings of texts. When there are many texts with sensitive information, the workload will be relatively large. Hassan et al. [3] proposed a more general solution to process the problem of text anonymization based on word embeddings, which has reported high recognition efficiency, but the model of this method needs to be trained by a large-scale corpus, and therefore the performance of the obtained model is unstable. Especially, the quality of the word vector of low-frequency words is not high possibly because there is only one-hot representation of a word and inevitably makes the representation of synonym become a problem. Neerbeky et al. [4] used a recurrent neural network (RNN) to assign sensitivity scores to the semantic components of each sentence structure for achieving the purpose of sensitive word detection with high accuracy, but the trained model of this method is not effective for processing sensitive words with multiple meanings.

The task of sensitive words recognition in this paper aims at recognizing sensitive words in the unstructured dialogue texts in electric power customer services. The goal of customer services systems is to receive the calls from electric power customers, answer and resolve some questions issued by customers. The conversations between electric power customers and the custom service staffs are often recorded and transformed into the dialogue texts stored by the electric power customer services system. For improving the efficiency of customer services, the historical dialogue texts need to be analyzed. However, the dialogue texts contain some privacy information of electric power customers such as the numbers of electric meters, customers' names, addresses, their ID cards information, bank accounts information, and other private information on their electricity usage. Before these texts are used for further analysis, the sensitive information mentioned above should be detected and masked by some character replacements for preventing them from leaking to the public.

However, the existing approaches for detecting and recognizing sensitive words are not efficient to process the dialogue texts in electric power customer services. The dialogue texts are often colloquial because electric power customers have no the trainings how to use more formal terms to communicate with custom service staffs. In the situation, customers often can say some OOV words that are close related to the professional terms (synonyms) but do not belong to the professional terms. For examples, electric power customers may say “personal identity” instead of “ID card number”, and say “meter value” instead of “scale of electric meter”, and so on. Most of these colloquial expressions do not contain the professional terms, so it is difficult for recognizing sensitive words to determine how close the relationship between the colloquial terms and the professional terms is. Unfortunately, the traditional approaches for sensitive words recognition mentioned above, including regulated expression techniques and learning based similarity computation, are inefficient to detect sensitive words in dialogue texts due to the fact that the colloquialism and synonyms exist in dialogue texts [7].

In this paper, we proposed an OOV based approach for recognizing sensitive words in the dialogue texts of electric power customer services. Firstly, the preprocessed text is trained by word vector and mapped to word vector matrix. Secondly, we combine the semantic similarity based on word embeddings and structural semantic similarity based on HowNet for recognizing sensitive OOV words in the dialogue texts. The similarity between text word vector and standard sensitive words is calculated by word vector similarity model and HowNet similarity model, respectively. Furthermore, the two kinds of similarities are empirically weighted and comprehensively obtain the overall similarity between OOV words and standard sensitive words. The related experiments were made, and the experimental results show that our method has higher recognition accuracy in comparison with the popular approaches.

The paper is organized as follows. Section 1 is the introduction. In Sect. 2, we give the framework of our approach. In Sect. 3, we respectively discuss the two kinds of semantic similarity models. Section 4 is to compute the overall similarity between standard sensitive words and OOV words. Section 5 is the experiment and analysis. Section 6 is the conclusion.

2 Sensitive Word Recognition Framework Based on Semantic Similarity

The framework proposed is mainly composed of three parts: text preprocessing, OOV words semantic similarity computation, and OOV words detection and desensitization, which is shown in Fig. 1.

Text preprocessing: In the process of Chinese text sensitive word recognition, the dialogue texts should be preprocessed first. In this article, the dialogue texts was segmented by the jiaba word segmentation tool. Because sensitive words can be composed of multiple words, they may be divided into multiple words, which destroys the meaning of the words themselves. Therefore, we add these sensitive words with clear characteristics to the word segmentation dictionary, so as to maintain the integrity of this kind of sensitive words in the word segmentation; At the same time, in order to save storage space and improve search efficiency, it is necessary to filter out the stop words such as function words and non-search words in the dialogue texts after word segmentation.

OOV words semantic similarity computation: To make the sensitive word recognition more accurate, we proposed an OOV based approach for recognizing sensitive words in the dialogue texts of electric power customer services which combining the semantic similarity based on word embeddings and structural semantic similarity based on HowNet. Firstly, we train large-scale word vector through Chinese Wikipedia corpus and existing text corpus. There are two training models for Word2vec word vector, CBOW model and SKIP Gram model [6, 8]. Here, we use the CROW model with context relationship to train the data [8], and express each word in the pre-processed text as a vector with appropriate dimensions [5]. Here, we set the dimension as 300 and map the text segmentation texts to the word vector matrix W.

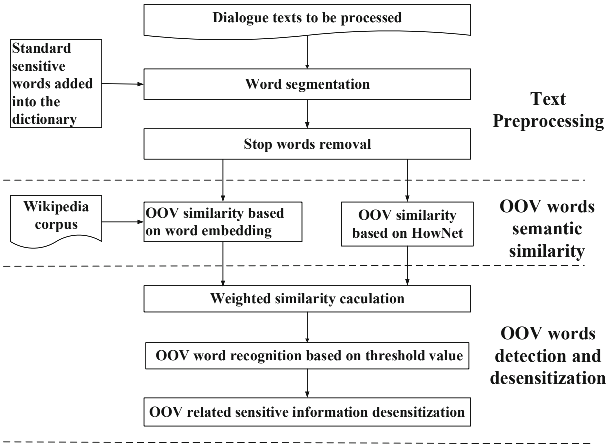
Dialogue texts to be processed
Standard sensitive words added into the dictionary
Word segmentation
Text Preprocessing
Stop words removal
Wikipedia corpus
OOV similarity based on word embedding
OOV similarity based on HowNet
OOV words semantic similarity
Weighted similarity caculation
OOV word recognition based on threshold value
OOV words detection and desensitization
OOV related sensitive information desensitization

Fig. 1. OOV word recognition framework based on semantic similarity.

Secondly, the word embeddings similarity model is used to calculate the similarity of word vectors and standard sensitive words in turn, also HowNet similarity model is used to calculate the similarity of word vectors and standard sensitive words in turn. Finally, we will obtain two kinds of similarities calculated by the two models.

OOV words detection and desensitization: The two kinds of similarities are empirically weighted and obtain the overall similarity between word vectors and standard sensitive words finally. All the words in the texts are judged whether they are OOV words through the set threshold, and if the similarity of one word and the standard sensitive word is greater than the set threshold, the word is OOV word. Then the sensitive information corresponding to the OOV word is found in the dialogue texts and the sensitive information will be masked.

3 Semantic Similarity Calculation Method

Using Word2vec characterized by vectorizing words so as to accurately measure the relationship between different words can effectively identify OOV words and improve recognition performance [10]. But this method needs to be trained by a large-scale corpus, and therefore the performance of the model is unstable, so here we added the model based on HowNet and combined the two models to identify OOV words.

3.1 Semantic Similarity Calculation Method Based on Word2vec

In Sect. 2, we mentioned that we need to map the text segmentation words to the word vector matrix W before we calculate the similarity between the word vectors in W and standard sensitive words. The similarity calculation methods of Word2vec mainly include Euclidean distance, Jaccard similarity, Cosine similarity and so on. In this paper, cosine similarity is used to calculate the similarity between word vectors in W and standard sensitive words. Cosine similarity refers to the cosine value between two vectors in vector space as the similarity of two vectors. The closer the cosine value is to 1, the smaller the included angle and the greater the similarity, that is, the more similar the two word vectors are. Assume that the word vector $w_k = x_{k1}x_{k2}...x_{kn}$ in W, and the standard sensitive word is $L = y_1y_2...y_n$. The equation for calculating the similarity S_{w_k} between w_k and standard sensitive words L is shown in Eq. (1).

$$S_{w_k} = \frac{\sum_{i=1}^{n}(x_{ki}y_i)}{\sqrt{\sum_{i=1}^{n}(x_{ki})^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{1}$$

where, x_{ki} represents the word vector obtained by word bag model for word w_k, and y_i represents the word vector obtained by standard sensitive word L through word bag model. S_{w_k} is the similarity calculated by word embeddings similarity model between the k^{th} word w_k and the standard sensitive word L. And here, $n = 300$.

3.2 Semantic Similarity Calculation Method Based on HotNet

HowNet was developed by human experts, where words are the smallest unit of use, and have exact meanings captured. Different meanings of words are defined as the concepts [9] represented by semantic expressions composed of several sememes [11]. The meanings expressed by these sememes are clear and fixed. There are ten sememe hierarchical trees such as event class, attribute class, entity class and attribute value class. There is no reachable path between sememes of different trees, and there is only one reachable path with length n between two different sememes in the same tree. The path length of these two sememes is the semantic distance between sememes. The method of semantic similarity calculation based on HowNet is used. Specifically, for the k^{th} word w_k in the preprocessed text, suppose that it has m concepts, namely, $s_{k1}, s_{k2}, \ldots, s_{km}$, and the standard sensitive word L is l_1, so the semantic similarity between word w_k and standard sensitive word L can be expressed as the maximum value of similarity between concepts:

$$sim(w_k, L) = \max_{i=1,2,\ldots,m} sim(s_{ki}, l_1) \tag{2}$$

where, $sim(w_k, L)$ is the similarity between the k^{th} word w_k and the standard sensitive word L, and $sim(s_{k1}, l_1)$ is the similarity between the concepts of w_k and L.

Equation (2) is to calculate the semantic similarity between concepts. Next, we calculate the similarity between sememes corresponding to concepts, that is, calculate the semantic distance of sememes. Assuming that the sememe of the concept s_{k1} of the word w_k is p_1, the sememe of the concept l_1 of the standard sensitive word L is p_2, and the distance of p_1 and p_2 in the hierarchical architecture is $\mathrm{dis}(p_1, p_2)$, the similarity between the two sememes is as follows:

$$sim(p_1, p_2) = \frac{\alpha}{\alpha + dis(p_1, p_2)} \tag{3}$$

where α is an adjustable parameter.

Semantic expressions of concepts can be divided into four parts: the first basic sememe, other basic sememes, relational sememes and symbolic sememes, so the overall similarity of the two concepts s_{k1} and l_1 is expressed as follows:

$$sim(s_{k1}, l_1) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} sim_j(s_{k1}, l_1) \tag{4}$$

where $sim(s_{k1}, l_1)$ is the similarity of class j sememe, and $\beta_i (1 \leq i \leq 4)$ is an adjustable parameter, which satisfies the following conditions: $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$, $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$. Therefore, the similarity $sim(w_k, L)$ of w_k and L is obtained by calculating the sum of the sememe similarities of two words, the similarity $sim(s_{k1}, l_1)$ of all the concepts of w_k and L is maximized, and finally the similarity S_{h_k} of w_k and L is obtained.

4 Recognition and Desensitization of OOV Words

4.1 Recognition of OOV Words

After the similarity of w_k and the standard sensitive word L is calculated by the two models respectively, the two kinds of similarities are weighted and we further obtains the overall similarity S_k of w_k and L. The specific definition is as shown in Eq. (5).

$$S_k = \alpha S_{w_k} + (1 - \alpha) S_{h_k} \tag{5}$$

where α is a threshold that needs to be set manually, and the importance of two similarity calculation methods in this model can be adjusted by α.

After calculating the similarity S_k between each word and standard sensitive words, we needs to set a threshold to determine whether the word is an OOV word. Assuming that the threshold $\beta = 0.8$, if the similarity S_k satisfies: $S_k \geq \beta$, then this word is an OOV word which we are looking for.

4.2 Desensitization of Sensitive Information

Sensitive information desensitization refers to the deformation of sensitive information through desensitization rules, so as to realize the reliable protection of sensitive private data. In this paper, the sensitive information is desensitized by hiding, that is, the sensitive information is masked by replacement *, and the sensitive information is covered. The specific process of desensitization is as follows: compare S_k with β in turn. If $S_k \geq \beta$, where $k = 1, 2, \ldots, m$, search the text for sensitive information corresponding to the OOV word w_k not far from the matching distance w_k. If the matching is successful, replace the sensitive information with *, and save the replaced text.

5 Experiment and Analysis

5.1 Datasets

The dataset used was obtained by transforming the collected dialogue voice of electric power customers and the customer services staffs into dialogue text over a period of three months, which contained a total of 2000 dialogue text data.

5.2 Evaluation Index

In order to verify the rationality and performance of the model of this system, Precision, Recall and F-measure are used as evaluation indexes as follows.

(1) Precision: $P = \frac{|SI \cap SC|}{|SI|}$

(2) Recall: $R = \frac{|SI|}{|SI \cup SC|}$

(3) F1-measure: $F_1 = \frac{2RP}{R+P}$

where SI is the set of sensitive words identified from all documents, and SC is the set of correct sensitive words contained in all documents. $|S|$ is to represent the cardinality of set S.

5.3 Method Statement

In order to verify the better performance of the method we proposed, the performance of this method is compared with the regulated expression technique and word embedding method. Regulated expression (RE) technique strictly matches the words in the texts with standard sensitive word. If this standard sensitive word is in the text, the recognition is successful. Word embedding (WE) converts a Word in the text into a vector and detects a word similar in meaning to a standard sensitive word. It can identify the synonyms of standard sensitive words.

5.4 Experiment and Result Analysis

Determination of Weight. There are two weights α and β. α refers to the importance of the similarity calculated by the word embedding similarity model in the final calculation. β refers to a set threshold. When the similarity is greater than the threshold, the similar word can be judged as an OOV word corresponding to the sensitive word. The α and β need to be determined manually. 500 text datasets were extracted to experiment on α and β respectively, the accuracy of α and β at different values was obtained. The experimental results are shown in Fig. 2 and Fig. 3.

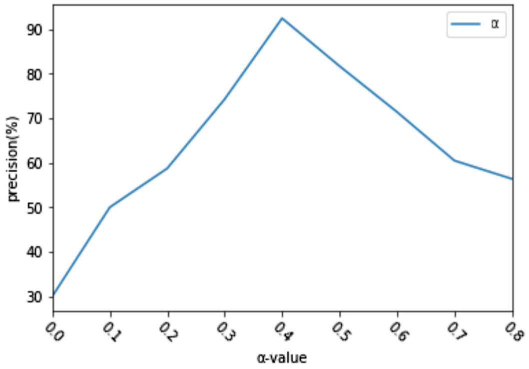

Fig. 2. The experimental results of α under different values

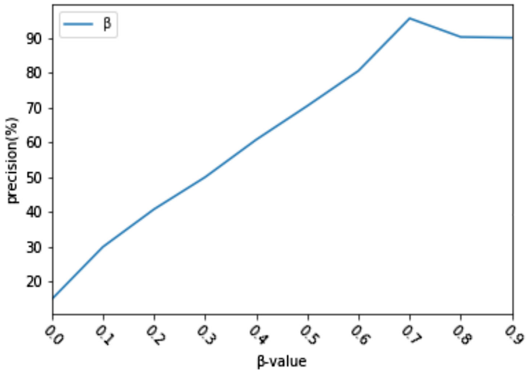

Fig. 3. The experimental results of β under different values

As shown in Fig. 2 and Fig. 3, for α, the sensitivity word recognition accuracy is the highest when $\alpha = 0.4$, so $\alpha = 0.4$; For β, the sensitivity word recognition accuracy is highest when $\beta = 0.7$, so the threshold value $\beta = 0.7$.

Performance Comparison and Analysis. The improved model is compared with the traditional model. The test results on the same dataset are shown in Fig. 4.

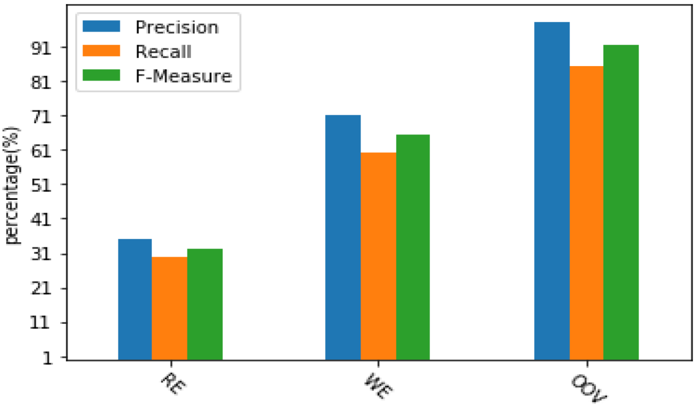

Fig. 4. Experimental comparison with word embedding and regulated expression

According to the results shown in Fig. 4, the values of precision, recall and F-measure of the OOV method we proposed are higher than the values of other two methods. The reason is probably because the RE method is only strictly matching the standard sensitive words, and consequently only strictly matching standard sensitive words can recognize the sensitive words from the text. The synonyms that are semantically close to sensitive words, i.e., the OOV words, cannot be recognized, and therefore the RE method has the lowest performance w.r.t the three indexes. In contrast, although the WE method can identify standard sensitive words and words similar in semantics to standard sensitive words, but the similarity calculation based on the WE method just relies on the statistics based word embedding that cannot exactly capture the true meanings of two words and further differentiate between them. Our OOV method obviously escapes from these limitations by combining the word embedding semantics and the structural semantics residing in HowNet. So the semantic similarity calculation based on the OOV approach will be more accurate and flexible than the other two methods.

6 Conclusion

We proposed an OOV based approach for recognizing sensitive words in the dialogue texts of electric power customer services. Our experiments show that our method has higher recognition accuracy in comparison with the popular approaches. Because in our method, we can not only match standard sensitive words precisely, but also detect the OOV words.

In the future work, we can study more efficient methods based on the approach we proposed in this paper.

Acknowledgements. This work is supported by State Grid Company Research Project "Research on Key Technology of Security of Use and Sharing of Power Grid Marketing Sensitive Data Based on Artificial Intelligence" (No. SGTYHT/19-JS-215).

References

1. Zhou, Y., Gao, C.: Research and improvement of a multi-pattern matching algorithm based on double hash. In: 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, pp. 1772–1776 (2017). https://doi.org/10.1109/compcomm
2. Yu, H., Zhang, X., Fu, C.: Research and application of change form of sensitive words recognition algorithm based on decision tree. Appl. Res. Comput. pp. 1–7 (2019)
3. Hassan, F., Sánchez, D., Soria-Comas, J., Domingo-Ferrer, J.: Automatic anonymization of textual documents: detecting sensitive information via word embeddings. In: Proceedings of 2019 18th IEEE International Conference On Trust, Security And Privacy, Rotorua, New Zealand, pp. 358–365 (2019)
4. Neerbeky, J., Assentz, I., Dolog, P.: TABOO: detecting unstructured sensitive information using recursive neural networks. In: IEEE 33rd International Conference on Data Engineering (ICDE), San Diego, CA, pp. 1399–1400 (2017) https://doi.org/10.1109/icde.2017.195
5. Chen, Y., Huang, S., Lee, H., Wang, Y., Shen, C.: Audio word2vec: sequence-to-sequence autoencoding for unsupervised learning of audio segmentation and representation. IEEE/ACM Trans. Audio, Speech, Lang. Process. **27**(9), 1481–1493 (2019). https://doi.org/10.1109/TASLP.2019.2922832
6. Ding, H., Yu, H., Qi, K.: Research on semantic prediction analysis of tibetan text based on word2Vec. J. Phys: Conf. Ser. **1187**(5), 52–58 (2019)
7. Zeng, J., Duan, J., Wu, C.: Adaptive topic modeling for detection objectionable text. In: Proceedings of 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence. Atlanta, pp. 381–388. IEEE (2013)
8. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
9. Yuan, X.: HowNet: research on the similarity calculation of Yiyuan. Liaoning Univ. Nat. Sci. Ed. **38**(4), 358–361 (2011)
10. Jin, G., Shi, Y., Wei, Z., Wang, Y., Liu, J.: A sensitive content recognition technology based on Word2vec. Commun. technol. **52**(11), 2750–2756 (2019)
11. Nie, H.M., Zhou, J.Q., Guo, Q., Huang, Z.Q.: Improved semantic similarity method based on HowNet for text clustering. In: 5th International Conference on Information Science and Control Engineering (ICISCE), pp. 266–269 (2018)