Extremely Short Chinese Text Classification Method Based on Bidirectional Semantic Extension

Yongzeng Yue, Yuhong Zhang, Xuegang Hu, Peipei Li

Key Laboratory of knowledge Engineering with Big Data; (Hefei University of Technology), Ministry of Education; School of Computer Science and Information Engineering; Hefei University of Technology, Hefei, 23061, China

ABSTRACT. Short text classification methods have achieved significant progress and wide application on text data such as Twitter and Weibo. However, the extremely short chinese texts like tax invoice data are different with traditional short texts in lackness of contextual semantic information, feature sparseness and extremely short length. The existing short text classification methods are difficult to achieve a satisfactory performance in these texts. To address these problems, this paper proposes a text classification method based on bidirectional semantic extension for extremely short texts like Chinese tax invoice data. More specifically, firstly, the Chinese knowledge graph is introduced for extending bidirectional semantic of texts and label data to expand the extremely short texts and ease the problem of feature sparseness; secondly, the hash vectorization is used to avoid the semantic problem caused by the lackness of contextual information. Experimental results conducted the real tax invoice dataset demonstrate the effectiveness of our proposed method.

1. INTRODUCTION

In recent years, various types of goods and services have emerged with the prosperity and development of the electronic commodity economy. The types of invoices that customers to purchase product and services are increasing. Due to the lack of systematic and professional tax invoice training, unstandardized and even incorrect tax invoice names often appeared, which bought great challenges to the classification of tax invoice data. How to effectively classify Chinese tax invoice data becomes one of the important tasks to be solved urgently. The length of the tax invoice name is very short, usually composed of several words or phrases, such as: Mold repair, Construction labor fee etc.

In the tax invoice classification task, the State Taxation Administration has issued a unified classification standard, namely the tax code record table which divides the product and services name into more than 4,200 categories; then each category is encoded. Finally, the corresponding product and service names in each category are added to the items in the category as keywords. Some examples are shown in Table 1:

| Keyword | Category |
|---------|----------|
| Mold repair | Tax classification 1 |
| Construction labor fee | Tax classification 2 |

It can be seen from Table 1 that the number of keywords in each class attribute are fewer, the feature information is sparse, and it is very difficult to cover the products and services in application.

The traditional short text classification problem has gained extensive attention [1-6]. However, most of existing methods focus on the classification of instant information such as Weibo, product reviews, etc., which can be mainly divided into two types, the first type is the method without external corpus. This type of method is classified by various machine learning algorithms using the characteristics of
short text itself. Yu et al. [7] proposed an open source short text classification method and analysis tool called LibShortText, which speeded up the training and testing of short texts to a certain extent; Wang et al. [8] proposed the NBSVM algorithm which utilized the naïve Bayesian log count ratio as the eigenvalue of the SVM; Niculae et al. [9] proposed a short text classification tool which called TextGrocery, it supports both Chinese and English corpora. However, the a force mentioned methods [7]-[9] have not solved the problem of feature sparseness in short text datasets. Wang et al. [10] utilized TF-IDF to calculate the weight of word vectors in short texts, and then Word2Vec vectorization are introduced for feature representation. However, Word2Vec vectorization needs the large number of training samples to reach a satisfied performance. Huang et al. [11] extracted the training categories from the chi-square statistics to reconstruct the training set, which improves the efficiency of the KNN classification algorithm, but fails to improve the classification accuracy.

Table 1. Commodity tax code record form

| Tax code               | Product and service name      | Keywords                                           |
|-----------------------|-------------------------------|----------------------------------------------------|
| 2020000000000000000   | Repair and replacement work   | Repair and replacement, Replacement, Repair, Replacement management |
| 3050400000000000000   | Decoration services           | Decoration Services, Decoration, Decorate          |
| 5020000000000000000   | Structures                    | The road, Bridge, Tunnel, Dam                     |

The second type is the method which introduced an external corpus for feature extension. Zelikovitz et al. [10] introduced the external corpus to determine the class label of the prediction data by judging the similarity between the training data and the prediction data, however, this method requires that the external corpus must be related to the experimental data; Fan et al. [11], Gabrilovich et al. [12] and Banerjee et al. [13] which leveraged the network knowledge base Wikipedia to extend the feature of short text for improving classification accuracy. However, although the information obtained from Wikipedia is the annotation content of the entry, there is no targeted expansion for feature information or class label; Wang et al. [14] employed the association rules between the training set and the test set to extend the feature words in the test set, however, the training data in this method need to reach a certain scale.

In the real Chinese tax invoice dataset, the length of the Chinese tax invoice name is much shorter than the traditional short texts such as Weibo and product reviews. The included feature information of these texts is sparse, and the association between the invoice name and the class label is weak, every tax invoice name is independent with each other and does not contain context information. To address these problems, this paper proposes an extremely short Chinese text classification method based on bidirectional semantic extension. This method utilized the Chinese knowledge map "word forest", and simultaneously queries the tax invoice name and the classification result of the class label to query the synonyms, expanding the text length and feature information of the short text, and the hash vectorization method avoids the semantic vectorization problem caused by the lack of context information. Finally, the experimental results on the real tax invoice dataset supported by the project show that the proposed method can significantly improve the classification accuracy.

2. RELATED WORK

2.1 Short Text Classification
Short text classification refers to the process of classifying short-length texts by means of machine learning and data mining. Different from traditional text processing methods, short text classification methods consider the issues of shorter text length and sparse features.

Ma et al. [6] proposed the short text classification method based on probabilistic semantic distribution. Firstly, they transforms the text into vector data by querying the vector dictionary. Then, the Gaussian
mixture is used to train the general background semantic model of the unmarked data, and the training data is used for a general purpose. The model adaptively obtains the semantic distribution model of each target domain, and finally calculates the probability that the short text belonging to the domain model, and obtains the final classification result.

Yu et al. [7] proposed the open-source text classification and analysis tool LibShortText based on linear kernel SVM includes three modules: converter, classifier and analyzer. Firstly, the part of speech is marked and deactivated. Then feature representation is selected and a linear SVM is used for classification finally. Compared with traditional text analysis and mining tools, the training efficiency and classification accuracy for short texts are significantly improved. Wang et al. [8] proposed the NBSVM algorithm used NB log count ratio as the eigenvalue of SVM to construct a linearity classifier. Niculae et al. proposed [9] the short text classification tool TextGrocery based on LibLinear [18] and Jieba participle [19] supports both Chinese and English corpora.

2.2 Chinese Knowledge Map

Many research have been conducted on the problem of collecting knowledge bases from online encyclopedias [20-22]. Although the knowledge bases compiled by these research institutes play an important role in the problem of text processing, most of them are in English, and there are fewer Chinese knowledge bases. Although many existing knowledge extraction systems can be used to build Chinese knowledge bases, they still face the following two problems: first, it takes a lot of man power to build ontology and supervised knowledge extraction models; second, the frequency is very slow in the future content update.

In order to address the above problems, the knowledge workshop team from the Graphic Data Management Laboratory of Fudan University (GDM@FUDAN) has developed a knowledge base system that can automatically generate an ever-expanding and constantly updated knowledge base system—CN-DBpedia [23]. The basic idea of the system is to first extract data from Baidu Encyclopedia, Interactive Encyclopedia and Chinese Wikipedia through crawlers, then standardize the attributes and values of the obtained data, and then the knowledge base by reusing the ontology in DBpedia [20] information is refined, and the crowdsourcing method is used finally to correct the information in the knowledge base through error detection and error correction. The knowledge base is kept fresh by proactive updates and regular updates. However, the entity relationship links provided by CN-DBpedia cannot accurately describe the name information of tax invoice. For example:

**Example 1:** In the “Construction labor fee”, the link entity corresponding to the word “Building” in CN-DBpedia is “General name for buildings and structures” and “Professional disciplines”; “Labor” does not have a corresponding link entity.

**Example 2:** In “Mold repair”, the link entity corresponding to the word “Mold” in CN-DBpedia is “Various molds and tools” and “Mold game platform”; “Maintenance” does not have a corresponding link entity.

It can be seen that the extension of CN-DBpedia is an annotative description of the entity, which causes semantic generalization and fails to solve the problem of lackness semantic information and feature sparseness faced by tax invoice data.

The essence of the online dictionary "word forest" is a semantic network composed of Chinese. It uses nodes and edges to express various entities, concepts and various semantic relationships between them. It is a data compiled and organized by the online dictionary team of "word forest" that has been working in the text industry for a long time. The data includes Chinese word interpretation, rest language, poetry dictionary, synonyms and antonyms, etc. The semantic network is created by methods such as algorithm aggregation, indexing and abstracting forms in an intelligent online dictionary.
3. EXTREMELY SHORT CHINESE TEXT CLASSIFICATION METHOD BASED ON
BIDIRECTIONAL SEMANTIC EXTENSION

3.1 Main Framework
Due to the short text length of the Chinese invoice name, the sparse data features, the lack of class label
attributes, and the lack of context, the association between each class and the corresponding label is very
weak. The common short text classification method is difficult to achieve satisfied performance on the
Chinese tax invoice data set. This paper proposes a bidirectional semantic extension based extremely
short Chinese text classification (BSE-ESTC).

The method firstly does word segmentation for the Chinese tax invoice data; then separately
classifies the word segmentation result and the class label attribute word through the online dictionary
“word forest” to query the synonym, the query result is saved and collected by the network crawler, and
then the found result is added to the word segmentation result. In the middle, the semantic expansion is
carried out to achieve the purpose of expanding the feature quantity of extremely short texts; then the
Hashing Vectorizer [24] is used to vectorize the segmentation results; finally, the experimental data is
divided into training sets and test sets. The training set data is used to train the Passive Aggressive
Classifier [25] of the incremental learning, and the test set data is used for the classification test. The
details of the specific semantic extension technology are detailed in the following sections.

3.2 Chinese Word Segmentation
First, we use the Jieba word segmentation tool to classify the Chinese invoice name in precise mode and
search engine mode. Take “Construction labor fee” as example 1, “Mold repair” as example 2. The result of the word segmentation of example 1 is:

The result of the word segmentation in the precise mode: Building, Labor; the name of the invoice
is divided into two words: “Building” and “Labor”;

The result of the word segmentation in the search engine mode: Building, Service, Labor; because
the search engine model is based on the precise model, the long word is further divided, that is, the word
is continued on the basis of the term “Labor”, and then cut out the term “Service”.

Example 2 has the same word segmentation results in the precise mode and the search engine mode,
which is:

Mold, Maintenance; the invoice name is divided into two words “Mold” and “Maintenance”.

3.3 Signed Hash Trick method
In the extremely short text classification of Chinese tax invoice data, due to the high sparsity of the text,
Weinberger et al. [26] proposed the problem of word frequency accumulation and feature value increase
we use the Signed Hash Trick method. In order to avoid the mapping of the two original features to the
same position after hashing, Signed Hash Trick is further improved based on the Hash Trick method. In
addition to the original hash function \( h(i) \), here is a hash function as shown in Eq. (1):

\[
\xi : N \rightarrow \pm 1
\]

\[
\tilde{\phi}(j) = \sum_{i \in T_M^{-j}} \xi(i) \cdot \phi(i)
\]

In Eq. (2), the post-hash feature is still an unbiased estimate and does not cause the value of some
hash positions to be too large. In the classification problem of Chinese tax invoice data, the high sparsity
of the text makes the hash map after dimensionality reduction can be well represented [27].

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental Data Set
The experimental data comes from the real Chinese tax invoice data set, which mainly includes two
parts of the product name and commodity code. The product name is the name of each type of invoice,
and the commodity code is the code in the commodity tax code table corresponding to the commodity name.

In our experiment, firstly, two groups are randomly selected in the Chinese tax invoice data set, each group contains 15000 instance as experimental data, and each set of extracted data is divided into training data and test data according to the ratio of 9:1. The proposed method compares the accuracy of the short text classification method with the short text classification method before and after the bidirectional semantic extension.

4.2 Comparison Method
Because the different modes in the Jieba word segmentation tool have a great influence on the accuracy of the classification, the proposed BSE-ESTC method is tested in the precise mode and search engine mode from the Jieba word segmentation tool, which are called BSE-ESTC1 and BSE-ESTC2 respectively. We compare the proposed method with the following methods:

1) LibShortText [7]: A short text classification method based on linear classifier for fast training and testing.
2) NBSVM [8]: Short text classification method with Bayesian logarithmic count ratio as the eigenvalue.
3) TextGrocery [9]: A short text classification framework based on the Jieba word segmentation module, which simplifies the formula based on LibShortText and removes the parser and parameter analysis module.
4) BSE-ESTC1: Extremely short text classification method, firstly uses the Jieba word segmentation precise mode for word segmentation, and then the "word forest" online dictionary for bidirectional semantic expansion is introduced, and finally hash vectorization and classification are conducted.
5) BSE-ESTC2: Extremely short text classification method, first uses the Jieba word segmentation tool search engine mode for word segmentation, then the "word forest" online dictionary for bidirectional semantic expansion is introduced, and finally hash vectorization and classification are conducted.

4.3 Experimental Results And Analysis
The results of the experiment show that the proposed method has better performance than the short text classification method, which shows the proposed BSE-ESTC is more suitable for the classification of extremely short texts like tax invoice data. From the experimental results, we can obtain the following observations:

- Due to the particularity of tax invoice data, keywords usually consist of one or several words, some data has no keyword, so the classification effect of NBSVM is not ideal. The proposed BSE-ESTC expands the word segmentation results and keywords at the same time, effectively solves the problem of extremely short text length and feature sparseness, and improves the classification accuracy.
- LibShortText is suitable for solving the classification regression problem of large-scale data and high-dimensional sparse features. Therefore, the effect of low-dimensional and small-volume tax invoice dataset is not satisfied in invoice tax datasets.
- TextGrocery introduces the Jieba word segmentation tool as a built-in default tokenizer based on LibShortText, which simplifies the formula and removes the parser and parameter analysis module. But it does not change the nature of LibShortText based on linear SVM. Therefore, it does achieve satisfied performance in the extremely short text data set like tax invoice data.

Figure 1 shows the accuracy of all comparison methods in the two data sets.
In summary, the experimental results show that the classification accuracy of the proposed method is significantly higher than other short text classification methods.

### 4.4 Experimental Method Analysis

In order to analyze the robustness of the short text classification method based on bidirectional semantic extension proposed in this paper, we further verify by experiments, the comparison method is as follows:

- **a) ESTC1**: Short text classification method, first the Jieba word segmentation tool’s precision mode for word segmentation is used, then vectorize and classify are conducted.
- **b) ESTC2**: Short text classification method, first the Jieba word segmentation tool’s search engine mode for word segmentation is used, then vectorize and classify are conducted.
- **c) CN-ESTC1**: Short text classification method, first the Jieba word segmentation tool’s precision mode for word segmentation is used, then CN-DBpedia for semantic extension is introduced, and finally vectorize and classify are conducted.
- **d) CN-ESTC2**: Short text classification method, first the Jieba word segmentation tool’s search engine model for word segmentation is used, then CN-DBpedia for semantic extension is introduced, and finally vectorize and classify are conducted.

On the extracted 30,000 data sets, the accuracy, recall rate and F1 value test were performed on the ESTC1, ESTC2, BSE-ESTC1 and BSE-ESTC2 methods. The specific results are shown in Figure 2:

It can be seen from Figure 4 that the proposed method is based on the two modes of Jieba word segmentation tool. Since the search engine mode is based on the precise mode, the long words are further divided and have more feature quantities. It is seen that the accuracy, recall and F1 values of the ESTC2 and BSE-ESTC2 methods in the search engine mode are higher than the ESTC1 and BSE-ESTC1 methods in the precise mode.

Under the same conditions, the accuracy, recall rate and F1 value of the CN-ESTC1, CN-ESTC2, BSE-ESTC1 and BSE-ESTC2 methods were tested. Experimental results are shown in Figure 3:

(a) ESTC1 & BSE-ESTC1  (b) ESTC2 & BSE-ESTC2

Figure 2. Experimental comparison between our methods using different modes of word segmentation
It can be seen that the methods CN-ESTC1 and CN-ESTC2 use the CN-DBpedia to expand the word segmentation results based on the methods ESTC1 and ESTC2, and the accuracy, recall rate and F1 value in the accurate mode and the search engine mode are not increased. Compared with the accuracy that without the expansion, it has slightly decreased. Because the content obtained by CN-DBpedia is the annotation of the words, the semantic generalization is incurred and there is no effectively expanding in the feature quantity.

The comparison test results on the two data sets show that the effectiveness and robustness of our proposed methods, which are more suitable for the classification of extremely short texts of Chinese tax invoice data.

5. CONCLUSION
The existing short text classification method is mainly for data such as Weibo or Twitter, and there is not much analysis and research on the classification problem of extremely short text such as tax invoice data. Combined with the actual classification of tax invoices, this paper proposes an extremely short Chinese text classification method based on bidirectional semantic extension, which simultaneously expands the word segmentation results and class label attribute words to improve the classification accuracy. However, the work of this paper does not fully take the time complexity problem into consideration. This will be one of our future research work.

REFERENCES
[1] Wang Y, Zhou Z, Jin S, et al. Comparisons and selections of features and classifiers for short text classification[C]//IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2017, 261(1): 012018.
[2] Wang P, Xu B, Xu J, et al. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification[J]. Neurocomputing, 2016, 174: 806-814.
[3] Zhou Y, Xu J, Cao J, et al. Hybrid attention networks for Chinese short text classification[J]. Computación y Sistemas, 2017, 21(4): 759-769.
[4] Ma H, Xing Y, Wang S, et al. Leveraging Term Co-occurrence Distance and Strong Classification Features for Short Text Feature Selection[C]//International Conference on Knowledge Science, Engineering and Management. Springer, Cham, 2017: 67-75.
[5] Sun B, Zhao P. Feature extension for Chinese short text classification based on topical N-Grams[C]//Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on. IEEE, 2017: 477-482.
[6] Ma C L, Yan Y H. Short text classification based on probabilistic semantic distribution. Acta Automatica Sinica, 2016, 42(11): 1711-1717.
[7] Yu H, Ho C, Juan Y, et al. Libshorttext: A library for short-text classification and analysis[J]. Rapport interne, Department of Computer Science, National Taiwan University. Software available at http://www.csie.ntu.edu.tw/cjlin/libshorttext, 2013.
[8] Wang S, Manning C D. Baselines and bigrams: Simple, good sentiment and topic classification[C]//Proceedings of the 50th Annual Meeting of the Association for
Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics, 2012: 90-94.

[9] Niculae V, Sun K, Chen X, et al. Cornell Belief and Sentiment System at TAC 2016[C]//TAC. 2016.

[10] Zelikovitz S, Hirsh H. Improving short text classification using unlabeled background knowledge to assess document similarity[C]//Proceedings of the seventeenth international conference on machine learning. 2000, 2000: 1183-1190.

[11] Fan Y J, Liu H L. Research on Chinese short text classification based on Wikipedia[J]. XIANDAI TUSHU QINGBAO JISHU, 2012, (3): 47-52.

[12] Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In:Proceedings of the 20th International Joint Conference on Articial Intelligence. San Francisco, USA: Morgan Kaufmann, 2007. 1606-1611.

[13] Banerjee S, Ramanathan K, Gupta A. Clustering short texts using wikipedia[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2007:787-788.

[14] Wang X W, Fan X H, Zhao J, Method for Chinese short text classification based on feature extension[J]. 2009, 29(3): 843-845.

[15] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. Machine learning, 1997, 29(2-3): 131-163.

[16] Bayes T, Bayes T. An essay towards solving a problem in the doctrine of chances[J]. Resonance, 2003, 8(4):80-88.

[17] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3):273-297.

[18] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification[J]. Journal of machine learning research, 2008, 9(Aug): 1871-1874.

[19] Sun J. ‘Jieba’Chinese word segmentation tool[J]. 2012.

[20] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A Nucleus for a Web of Open Data. [C]// The Semantic Web, International Semantic Web Conference, Asian Semantic Web Conference, ISWC 2007 + Aswc 2007, Busan, Korea, November. DBLP, 2007:722-735.

[21] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]// Sigmoid Conference. 2008.

[22] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]// International Conference on World Wide Web. OAI, 2007.

[23] Xu B, Xu Y, Liang J, et al. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System[J]. 2017:428-438. Hearst M A, Dumais S T, Osuna E, et al. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4): 18-28.

[24] Pugh W. The Omega test: a fast and practical integer programming algorithm for dependence analysis[M]. 1991.

[25] Crammer K, Dekel O, Keshet J, et al. Online passive-aggressive algorithms[J]. Journal of Machine Learning Research, 2006, 7(Mar): 551-585.

[26] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg, “Feature hashing for large scale multitask learning,” in International Conference on Machine Learning, 2009, pp. 1113–1120.

[27] J. Attenberg, K. Weinberger, A. Dasgupta, A. Smola, and M. Zinkevich, “Collaborative email-spam filtering with the hashing trick,” in Proceedings of the Sixth Conference on Email and Anti-Spam, 2009.