Tibetan-Chinese cross-lingual word embeddings based on MUSE

Wei Ma 1*, Hongzhi Yu 1, Kun Zhao 1, Deshun Zhao 1, Jun Yang 1

1 Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu, 730030, P.R. China

*Corresponding author’s e-mail: mw1231@qq.com

Abstract. The idea of word embedding is based on the semantic distribution hypothesis of linguist Harris (1954), who believes that words with the same semantics are distributed in similar contexts. The learning of word embedding is a crucial technology in natural language processing. In recent years, cross-language word vectors have received more and more attention. Cross-language Word vectors can transfer knowledge between different languages. Most importantly, this transfer can occur between rich-resource and low-resource languages. This paper uses the Tibetan-Chinese Wikipedia corpus to train monolingual word vectors. Based on the Tibetan-Chinese bilingual translations, we use the supervised method in the MUSE library to train the Tibetan-Chinese bilingual cross-language word embeddings. In the experiment, we evaluate the result of word representation on the standard lexical semantic evaluation task. The results show that the method has a certain improvement in the semantic representation of the word embedding.

1. Introduction

Word Embedding is a method of obtaining a word-level vector representation (word vector) using distributed representation in the field of natural language processing, using co-occurring statistics from a large amount of text. The classic model is the NNLM proposed by Bengio et al. [1] in 2003, which obtained the word vector in the process of training the language model. In recent years, the monolingual Word Embedding technique has inspired the imagination of researchers in the field of natural language processing. Word vectors have been widely used in natural language processing tasks such as named entity recognition (Guo et al., 2014[2]), sentiment analysis (Socher et al., 2013[3]), and text categorization.

The idea of Word Embedding is based on the semantic distribution hypothesis of the linguist Harris (1954)[4], who believes that words of the same semantics are distributed in similar contexts. The current text representation methods are mainly Bag-of-Words (BOW) and distributed representation methods. The typical representative of distributed representation is Word Embedding. Word Embedding was first proposed by Hinton[5]in 1986. The word vector is a dense and low-dimensional real-value vector, and each dimension is represented by a real number. It also represents semantic and grammatical features, so that words with certain semantic relations are closer in the mathematical sense. This method better solves the dimensionality disaster problem existing in the traditional One-hot method, and also incorporates the semantic relationship between words into the representation of the text. The Word2Vec model is a word vector training model proposed by Mikolov et al. [6]in 2013. It is used to train learning words and phrase vectors, and the processing of text content is transformed into vector operations in vector space. The semantic similarity of the text is represented by the similarity in the vector space. Later, researchers extended the word vector representation to a phrase-level or sentence-level
representation. Global vectors (GloVe) (Pennington et al., 2014[7]) allows us to learn word representations via matrix factorization.

In recent years, cross-lingual vocabulary embedding has received more and more attention, mainly because: First, they support cross-lingual and multi-language semantics, that is, reasoning meanings in multi-linguistic contexts. It also calculates cross-lingual lexical similarities associated with many tasks, as well as many other applications, such as bilingual dictionary induction or cross-lingual information retrieval. Second, cross-lingual lexical embedding enables knowledge transfer between languages, most importantly between languages with rich resources and low resources. For languages with fewer resources such as Tibetan, the training data is less, and the hardware required for training is high. How to use other languages that already have abundant resources to train more expressive word vectors economically and efficiently is a research focus. For cross-lingual word vectors, Facebook opened the Multilingual Unsupervised or Supervised word Embedding (MUSE) [8] library in 2018. The library can use two methods to construct cross-language word vectors.

The Tibetan language is a low-resource language, and bilingual parallel corpus with other languages is particularly lacking. Tibetan language research has made great progress, but its development is relatively lagging behind other resources-rich and widely used languages. In the case of the Tibetan language, the study of Tibetan text processing is a challenging task. In languages such as English, spaces are used as natural delimiters between words, and as Chinese, there is no obvious distinction between Tibetan words. In this paper, based on the fastText word vector training method[9], we used the Tibetan-Chinese Wikipedia corpus to train the monolingual word vector. We used five dictionaries including the Tibetan-Chinese Dictionary to build a Tibetan-Chinese bilingual dictionary. The dictionary constructed 123,168 bilingual word pairs, of which 98,534 were in the training set dictionary and 24,634 in the test set dictionary.

According to the method of MUSE, two monolingual word vectors are analyzed by MUSE to obtain Tibetan Chinese cross-language word vectors and evaluated using standard lexical semantic evaluation tasks. The results show that this method has a certain improvement in the semantic representation of the single-word vector.

The rest part of the paper is organized as follows, the second part is a brief review of the previous cross-lingual word vector study. The third part is a new way of Tibetan-Chinese cross-lingual word vector training. The fourth part gives the experimental results, in the fifth part, we draw conclusions and outline future work.

2. Related Work
In recent years, more and more researchers have turned their attention to the related research of cross-lingual word vectors for two reasons. First, cross-lingual word vectors can infer the semantics of words in a multi-language environment. Second, cross-lingual word vectors can implement knowledge transfer between different languages. Mikolov, Le, and Sutskever (2013) [10] notice that the geometric relations that hold between words are similar across languages. Dinu, Lazaridou, and Baroni (2015) [11] discover that using MSE as the sub-objective for learning a projection matrix leads to the issue of hubness: some words tend to appear as nearest neighbors of many other words. They proposed a globally corrected neighbor retrieval method to overcome this issue. Faruqui and Dyer (2014) proposed a new CCA method for the first time: using canonical correlation analysis (CCA) to project two languages (or multiple) to new shared words. Embedded in the space. CCA learns a transfer matrix for each language so that all languages are transferred to a new space. Xiao and Guo (2014) [12] proposed to mix the corpus of multiple languages and directly train the new embedding space in the mixed corpus of the two languages. The main idea of MUSE is to use a linear mapping matrix W to project the source language’s word vectors into the target language’s word vectors. MUSE is a Python library for multilingual word embeddings, whose goal is to provide the community with state-of-the-art multilingual word embeddings and large-scale high-quality bilingual dictionaries for training and evaluation, include two methods, one supervised that uses a bilingual dictionary or identical character strings, and one unsupervised that does not use any parallel data[13].
This paper attempts to apply the MUSE to the construction of Tibetan-Chinese bilingual word vectors and tests the performance of bilingual word vectors in similarity tasks. Our contribution is to construct a bilingual word pair and try to construct the Tibetan-Chinese bilingual word vector space by MUSE. Experiments prove that the bilingual word vector we construct can significantly improve the semantic representation ability of the word vector.

3. Method

3.1. fastText
The fastText algorithm is a supervised model, similar to the CBOW architecture of word2vec. CBOW predicts intermediate words through context, while fastText predicts tags through context (this tag is the type of text, which is determined by manual annotation). From the model architecture, like CBOW, the fastText model also has three layers: input layer, hidden layer, output layer (Hierarchical Softmax, input is a number of words and their n-gram features, these features are used to represent a single document, the hidden layer is the superimposed average of multiple feature vectors, The hidden layer solves the maximum likelihood function, then constructs a Huffman tree according to the weights and model parameters of each category, and uses the Huffman tree as the output. If you use normal Softmax training, each label needs to be calculated, With the Huffman tree, the number of tags is large, the weight is high, and the Huffman coding is naturally shorter so that calculating the tags according to the Huffman coding path can greatly reduce the amount of calculation.

3.2. MUSE
Facebook's open-source Multilingual Unsupervised or Supervised Word Embedding (MUSE) is an unsupervised and supervised multilingual word embedded Python library that aligns bilingual word vector embedding spaces in an unsupervised or supervised manner. The supervisory method uses a bilingual dictionary. The unsupervised method does not use any parallel data, it establishes a bilingual dictionary between two languages by aligning the word embedding space in an unsupervised manner. Based on fastText, MUSE has the most advanced multi-language word embedding function in more than 30 languages. fastText is a library for efficient learning of word representations and text categorization. MUSE also has 110 large-scale, high-quality, real-life bilingual dictionaries to address the development and evaluation of cross-language word embedding and multilingual NLP methods.

3.3. Word Representation Evaluation
For the evaluation of the quality of the word vector, the most common and fastest evaluation method in the industry is to calculate the word similarity task and the word analogy task, although the quality of word vectors also needs to be applied to specific tasks for evaluation, including sentence classification, text classification[13], part-of-speech tagging[14], named entity recognition (NER)[15], etc., but these two tasks are still the most basic.

Word similarity task.

We evaluate the quality of our Chinese word vector representations on two different benchmarks that have been widely used to measure word similarity. The first one is the Wordsim-240[16] dataset containing 240 pairs of Chinese words that have been assigned similarity ratings by humans. The second benchmark is the Wordsim-296 dataset containing 296 pairs of Chinese words. For the Tibetan word vector, we made two Tibetan word vector evaluation data according to the Chinese evaluation data set, which are boWordsim-240 with 240 pairs of Tibetan words. BoWordsim-296 contains 296 pairs of Tibetan words. These data are obtained by means of machine translation plus manual verification.

In the word similarity task, Spearman's rank correlation coefficient ($\rho$) is generally used as the evaluation index, abbreviated as rho, which is an index to measure the dependence of two variables. It is an indicator of the dependence of two variables, which uses a monotonic equation to evaluate the correlation of two statistical variables.
We calculate the similarity between a given pair of words by the cosine similarity between their respective vector representation. We then report Spearman’s rank correlation coefficient (Myers and Well, 1995[17]) between the rankings produced by our model against the human rankings.

4. Experiment
In this section, we present our experimental results and perform some analyses to better understand our models.

4.1. Dataset
Training Data.
We use Wikipedia text as a dataset for training word vectors. Wikipedia is the largest free online encyclopedia, available in more than 200 different languages. Because the articles are curated, the corresponding text is high quality, making Wikipedia a great resource for (multilingual) natural language processing. It has been applied to many different tasks, such as information extraction, word sense disambiguation, multilingual entity linking. We downloaded the XML Wikipedia dumps from September 20, 2019. The first preprocessing step is to extract the text content from the XML dumps. Wikipedia has a lot of document parsing mature tools (such as gensim, Wikipedia extractor, etc.), we use the open-source tool Wikipedia extractor to complete the extraction of the text. The documents in the Wiki corpus contain traditional Chinese. We use the open-source toolkit Opencc to convert traditional Chinese into simplified ones. Now there are many English words (also some Japanese, German, etc.) in the corpus, in order to avoid affecting the effect of the trained word vector, we delete the English and empty brackets. Extracted Chinese corpus We chose to use the Jieba word segmentation tool for word segmentation to obtain the Chinese Wikipedia corpus to be trained. As the process of pre-processing with the Chinese Wikipedia corpus, Tibetan corpus also needs word segmentation. We chose the TIP-LAS open-source tool developed by Yachao Li [18] to segment the Tibetan text and obtain the Tibetan Wikipedia corpus to be trained. Translation pairs for MUSE.

For MUSE, two precise bilingual word pairs are required. We obtain the Tibetan-Chinese translation word pairs in two ways. First, we obtain the one-to-one bilingual entry from the Wikipedia open language link database file. We compare the Tibetan and Han-Tibetan language link data files and obtain nearly 5,000 bilingual words. The second way is that we extracted the noun-based translation word pairs from the existing Tibetan-Chinese dictionary. In total, more than 123168 translated words were obtained in two ways.

4.2. Results
In this section, we present the results of the experiments conducted to demonstrate the effectiveness of the proposed method with two groups of the experiment.

Table 1 shows the Spearman's correlation ratio obtained by calculating the similarity between two given word vectors. The first and fifth rows in the table show baseline scores obtained by using only Chinese and Tibetan monolingual vectors and others correspond to the performance of Tibetan-Chinese bilingual word vectors in different dimensions. For both similarity tasks, we have improved scores over the baseline. These results also illustrate that multilingual word vectors can help improve the distributed representation of word vectors.
5. Conclusion
In this paper, the MUSE is used to map the Tibetan and Chinese monolingual vectors to the same vector space, which realizes the construction of the Tibetan-Chinese bilingual word vector space. The motivation is that the cross-lingual word vector is the basis for realizing cross-lingual entity links. Resource-rich languages provide more resources for low-resource languages, enriching the language resources for the target language.

The training of Tibetan and Chinese language vector vectors is based on Wikipedia data, trained by the fastText method, and MUSE is used to map monolingual vectors to the same vector space. The results show that the Tibetan-Chinese bilingual word vector has a certain improvement in the semantic similarity evaluation score. However, this is only a basic task in the word vector evaluation task. In the future, we will also apply the bilingual word vector to the cross-language entity link, cross-language named entity recognition, cross-language information retrieval, and other tasks.

Acknowledgments
This research has been supported by the Fundamental Research Funds for the Central Universities, Northwest Minzu University (31920160005) and (31920190094), National Science and Technology Major Project(2017YFB1002103). Key Laboratory of China's Ethnic Languages and Information Technology (Northwest Minzu University), Ministry of Education. The authors gratefully acknowledge the financial support from Northwest Minzu University.

References
[1] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb): 1137-1155.
[2] Guo J, Che W, Wang H, et al. Revisiting embedding features for simple semi-supervised learning[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 110-120.
[3] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 1631-1642.
[4] Harris Z S. Distributional structure[J]. Word, 1954, 10(2-3): 146-162.
[5] Hinton G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science society. 1986, 1: 12.
[6] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
[7] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.
[8] A. Conneau*, G. Lample*, L. Denoyer, MA. Ranzato, H. Jégou, Word Translation Without Parallel Data

[9] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the ACL, 5, 135–146.

[10] Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation[J]. arXiv preprint arXiv:1309.4168, 2013.

[11] Dinu G, Lazaridou A, Baroni M. Improving zero-shot learning by mitigating the hubness problem[J]. arXiv preprint arXiv:1412.6568, 2014.

[12] Xiao M, Guo Y. Distributed word representation learning for cross-lingual dependency parsing[C]//Proceedings of the Eighteenth Conference on Computational Natural Language Learning. 2014: 119-129.

[13] G. Lample, A. Conneau, L. Denoyer, MA. Ranzato Unsupervised Machine Translation With Monolingual Data Only

[14] Tao Jiang, Yugang Dai, Ailin Li, and Hongzhi Yu. Tibetan Text Classification Using SVM and NB. In Proceeding of the 2nd National Conference on Information Technology and Computer Science.1160-11 65,2015.3.

[15] Tsai C T, Mayhew S, Roth D. Cross-lingual named entity recognition via wikification[C]//Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. 2016: 219-228.

[16] Su T R, Lee H Y. Learning Chinese word representations from glyphs of characters[J]. arXiv preprint arXiv:1708.04755, 2017.

[17] Well AD, Myers J L. Research design & statistical analysis[M]. Psychology Press, 2003.

[18] Li Y, Jiang J, Jia Y J, et al. TIP-LAS: an open-source toolkit for Tibetan word segmentation and POS tagging[J]. Computer Journal of Chinese Information, 2015, 29(15): 203-207.