Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230110

Word alignment of Chinese poetry parallel corpus based on word embedding technology

Xizhe Wang

Lanzhou University, Gansu, 730000, China

wangxzh20@lzu.edu.cn

Abstract. The research of word alignment technology provides the basis for many fields in natural language processing, such as speech recognition, bilingual dictionary writing, corpus construction and information retrieval. Based on the existing research foundation of parallel corpus for poetry sentence alignment, this paper studies how to align English-Chinese words. Firstly, the parallel corpus of Chinese poetry is preprocessed and translated into words, and the English poetry word vector is trained based on word embedding method, and the optimal corresponding word is judged by similarity calculation, and finally getting word alignment from parallel corpus sentence alignment.

Keywords: natural language processing, word alignment, word embedding, classical Chinese poetry.

1. Introduction

The bilingual parallel corpus is composed of two translated sentences. Bilingual parallel corpora not only play an important role in the training of machine translation models, but also have strong applications in bilingual lexicography, terminology research and application, cross-language information retrieval and so on[1]. There are many types of alignment in bilingual corpus, including discourse alignment, paragraph alignment, sentence alignment and word alignment. Among them, discourse alignment is easy.

The task of word alignment is to find the correspondence existing between words in parallel sentence pairs. The concept of word alignment was first introduced in the 1990s by Peter Brown at IBM Watson Research as an intermediate in statistical machine translation models. At that time, the translation quality of statistical machine translation was far superior to that of rule-based translation models, so word alignment received the attention of many scholars[2]. With the continuous development of machine translation, word alignment also plays an important role in neural machine translation. For example, they can inject external vocabulary into the inference process to improve the accuracy in the translation of domain-specific terms and low-frequency content words. Nowadays, the most important application of word alignment is to perform translation between annotations in the source language and annotations in the target language. For parallel corpus word alignment technology, the existing methods mainly include statistical machine translation based method, dictionary based method and language feature based method. Brown et al. proposed the IBM1-IBM5 model, which is the first statistical word alignment model proposed[3]. These five models were obtained under different assumptions. Brown

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/8/20230110

demonstrated and elaborated these five translation models in detail, laying a theoretical foundation for translation models. Vogel proposed a word alignment Model based on Hidden Markov Model [4]. Och et al. implemented and open-source the first statisc-based unsupervised word alignment tool, GIZA, based on the IBM word alignment model proposed by Brown[5]. In the source language sentence of the model, if the position of two words is closer, then the position of the aligned target word in the target language sentence is also closer, that is, the word alignment probability is related to the alignment position of the previous word.

For the word alignment of Chinese poetry parallel corpus, the existing research on Chinese poetry bilingual only focuses on the sentence alignment level, and the more fine-grained scope alignment research such as word alignment is still in the blank. The main work of this paper is to study how to align English-Chinese words based on the existing parallel corpora for the alignment of Tang poems and Song words and sentences. In terms of practical application, if the Chinese poetry learners or researchers find that the translation sentence contains unfamiliar terms, it may become a tricky problem to understand the terms. At this time, the bilingual parallel corpus can assist people to refer to the corresponding terms for translation.

2. Data preprocessing

2.1. Data cleaning

The purpose of data cleaning is twofold: the first is to make the data usable through cleaning; the second is to make the data more suitable for subsequent analysis work. The initial processing of the parallel corpus data is carried out by the following methods.

(1) removing redundant gaps, word separation, and lexical annotation
(2) Uniform character set, using UTF-8 as the uniform character set
(3) Normalization of English word forms
(4) convert all Chinese to simplified Chinese, and English uppercase to lowercase.
(5) Remove invalid characters, including Chinese and English punctuation, and special symbols.

2.2. Chinese poetry word segmentation

Word segmentation based on Chinese parts of ancient poems is the basis of word alignment work, and only after ensuring accurate results of word segmentation can follow-up research be carried out. Word segmentation is the process of recombining successive word sequences into semantically independent word sequences according to certain specifications. In English, Spaces are used as natural delimiters between words, while in Chinese, only words, sentences and paragraphs can be delimited simply by obvious delimiters, but words do not have a formal delimiter. Therefore, if word processing and analysis are to be carried out in Chinese text, word segmentation and extraction need to be carried out first, that is, word segmentation. The representative methods of Chinese word segmentation include shortest path segmentation, n-grammar segmentation, word construction by word segmentation, recurrent neural network segmentation, Transformer segmentation, etc. Due to the complexity of the grammar of the language, it is not easy to achieve word separation by grammar or word construction alone. We choose to use the Jiayan grammatical word splitting tool, which uses unsupervised, lexicon-free n-grammar and character-level hidden Markov model for automatic word splitting in ancient Chinese, and the tested results are better than the common word splitting tools such as jieba and Hanlp.

3. Word embedding and alignment

Word embedding technology is to represent text data in the form of numerical vectors that can be recognized by computers, which is the basis for text data to be processed by computers. The word vector can be regarded as the feature vector of the word, which is the embedding of the high-dimensional space with the dimension of the dictionary into a low-dimensional continuous vector space. The technology of mapping words to vectors in the real number domain is also called word embedding [6]. In natural language processing tasks, word embedding is a way to represent a word in a natural language by

DOI: 10.54254/2755-2721/8/20230110

representing each word as a point in an n-dimensional space, that is, a vector in a high-dimensional space. By this method, natural language computation can be transformed into vector computation. In order to achieve word alignment correspondence, the word2vec model is used to calculate the similarity between words.

The Word2vec model was proposed by the Google team in 2013. Its core idea is to predict the context of a keyword based on its context or predict the keyword based on the context of the keyword [7,8]. The Word2vec model consists of two structures: the continuous bag-of-words model (CBOW) and the skip-gram model, where CBOW infers the central word from the word vector of the context and skip-gram infers the context from the central word. Skip-gram is a neural network with a 3-layer structure: input layer, hidden layer, and output layer. skip-gram takes the current word as the input of the log-linear classifier and outputs other words within a certain range of the current word, i.e., the model inferred the original utterance from the target word, increasing the prediction range of the current word, which can improve the quality of the word vector, but it also causes the problem of increasing the computational complexity [9]. The input layer receives a one hot tensor as the input of the network, which stores the one-hot representation of the current sentence head.$V \in R^{1\times vocab_size}$The hidden layer multiplies the tensor V by a tensor, and uses the result as the output of the hidden layer to obtain a tensor of shape, which stores the word vector of the current sentence head; $W^1 \in R^{vocab_size\times embed_size}V \in R^{1\times embed_size}$The output layer multiplies the result of the hidden layer by another tensor to obtain a tensor of shape.$W^2 \in R^{embed_size\times vocab_size}V \in R^{1\times vocab_size}$After the softmax transformation of this tensor, the prediction result of the context using the current head word is obtained. Based on the results we can then train the word vector model. The preprocessed English poetry data set corpus is trained with skip-gram word vector. The ID of each word in the corpus is constructed according to the frequency of each word in the corpus, the higher the frequency, the smaller the ID, which is convenient for the management of the dictionary. Replacing each word with the corresponding ID is convenient for the neural network to process. Secondary sampling method is used to process the data and reduce the frequency of high-frequency words in the corpus The reduction method is to randomly discard high-frequency words, the higher the frequency, the higher the probability of being abandoned, and the lower the frequency, the lower the probability of being abandoned, so as to optimize the word vector training effect of the whole vocabulary. A sliding window is used to construct the training data to scan the corpus from left to right, and within each window, the central word predicts its context and forms the training data, which are all assembled into a mini-batch and fed into the network for training.

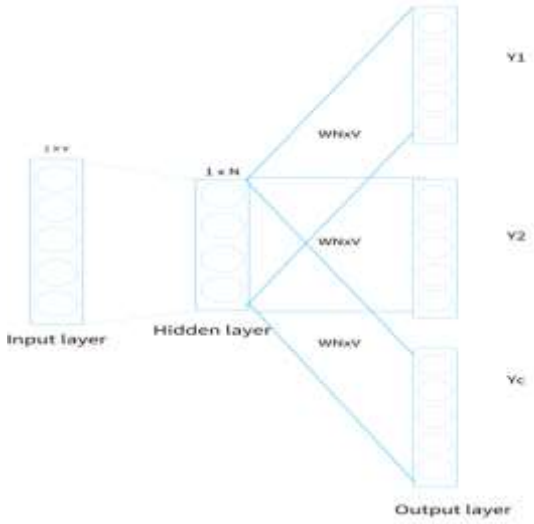

Figure 1. Skip-gram algorithm.

Accurate translation of Chinese poems is also indispensable to get accurate alignment results. After testing and comparing the existing common translation APIs (such as Baidu, Youdao, etc.) calls, Youdao translation API has higher accuracy in translating literary words. Finally, using a parallel corpus of

DOI: 10.54254/2755-2721/8/20230110

poems as a unit collection, we use cosine similarity to calculate the similarity between the English translation of Chinese words and the English corpus to determine whether there is word alignment between the two words. If the similarity is extremely high or much higher than other words in the same group, the two words are judged to be word aligned.

4. Analysis of results

After the alignment is completed, five poems are selected to test the accuracy of word alignment. Due to the differences in reading habits and grammar between Chinese and English, some words do not have strict word alignment, so they are marked with '/' when counting the results.

Taking Wen Tianxiang's SAILING ON LONELY OCEAN " as an example, the inspection examples and inspection results are as follows:

Example validation:

Table 1. Example of word alignment matching for "SAILING ON LONELY OCEAN".

Chinese words	Match the word pairs with the highest similarity	Highest cosine similarity	Correct word alignment of the original corpus	Is the word alignment accurate
Xinku	hard-delving	0.926179	hardship	no
zaofeng	Encounter-though	0.891965	though	yes
qi	rise-rose	0.939051	rose	yes
yijing	once-delving	0.943450	book	no
gange	war-fight	0.926031	fight	yes
liaoluo	sparse -desperately	0.927352	desperately	yes
sizhou	around-foe	0.889179	/	/
xing	star-year	0.929364	/	/
shanhe	land-land	1.000000	land	yes
posui	broken-desolate	0.893766	desolate	yes
feng	wind-down	0.898868	/	/
piao	flocculant-/	No match	/	no
shenshi	experience-or	0.857639	/	/
fuchen	sink-sink	1.000000	sink	yes
yu	rain-rain	1.000000	rain	yes
ping	duckweed-duckweed	1.000000	duckweed	yes
huangkong	fear-perilous	0.932842	perilous	yes
tantou	beach	1.000000	beach	yes
shuo	say-have	0.917459	/	/
lingding	lonely	1.000000	lonely	yes
yang	ocean	1.000000	ocean	yes
tan	sign-feel	0.932537	/	/
ren	man-man	1.000000	man	yes
sheng	grow-since	0.939532	/	/
zigu	since-since	1.000000	since olden days	no
shei	who-never	0.921427	/	/
wu	nothing-never	0.936880	never	yes
si	die-die	1.000000	die	yes
liu	leave-leave	1.000000	leave	yes
qu	take-leave	0.936982	leave	yes
danxin	loyalty- loyalist	0.961494	loyalist	yes
zhao	license-loyalty	0.908041	/	/
hanqing	history-history	1.000000	history	yes

Example analysis: A total of 33 words have been segmented in " SAILING ON LONELY OCEAN ", among which 24 words have corresponding relationships between Chinese and English bilingual corpora after testing, 20 words are aligned correctly, and 4 words have errors, including 2 errors caused by word segmentation errors and 2 errors caused by algorithm deviation.

DOI: 10.54254/2755-2721/8/20230110

Test results:

Table 2. Results of word alignment accuracy test for 5 poems.

Name of poetry	Total number of Chinese words	There exists word alignment word count	Accurately align the number of words	Alignment accuracy (exact words/existing words)
SAILING ON LONELY OCEAN	33	24	20	83.33%
FAREWELL TO A LUTIST	21	19	15	78.95%
MOORING AT GUAZHOU ISLE	22	18	16	88.89%
THE BOOK	20	16	14	87.5%
RETURN TO NATURE	30	28	23	78.57%
Total	126	105	90	84.76%

The test set achieves 84.76% accuracy.

5. Conclusion

This paper mainly studies how to align English and Chinese words based on the existing research on the parallel corpus of poem sentence alignment. According to the results of this experiment and test, the research of parallel corpus sentence alignment of ancient poetry word alignment has achieved a relatively good accuracy result. However, for the translation of more literary statements, free translation statements, even manual judgment, cannot achieve accurate alignment. Therefore, the follow-up work of the text can add human guidance to interfere with the alignment of sentences and explore the selection of user intervention points. Further improving the accuracy of high score words and expanding the amount of word embedding training data can be the research direction of improving the accuracy in the future.

References

[1] Huang Jiayue, Xiong Deyi. A Survey on Sentence Alignment.,2021,35(08):16-27.
[2] Brown P F, Cocke J, Della Pietra S A, et al. A statistical approach to machine translation[J]. Computational linguistics, 1990, 16(2): 79-85.
[3] Schoenemann, T. (2013). Training Nondeficient Variants of IBM-3 and IBM-4 for Word Alignment. ACL.
[4] Vogel S , Ney H , Tillmann C . HMM-based word alignment in statistical translation[C]// Proceedings of the 16th conference on Computational linguistics - Volume 2. Association for Computational Linguistics Morristown, 1996.
[5] Al-Onaizan, Y. , Curin, J. , Jahr, M. , Knight, K. , & Yarowsky, D. . (1999). Statistical machine translation: final report. john hopkins university center for.
[6] Chen Meng, He Zhiqiang, Wang Mengxue. Word embedded model research review [J]. Journal of hebei academy of sciences, 2021, 38 (02) : 8-16. DOI: 10.16191 / j.carol carroll nki. HBKX. 2021.02.002.
[7] Tomas Mikolov et al. Efficient Estimation of Word Representations in Vector Space[J]. CoRR, 2013, abs/1301.3781
[8] Tomas Mikolov et al. Distributed Representations of Words and Phrases and their Compositionality.[J]. CoRR, 2013, Abs / 1310.4546; 3111-3119
[9] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems NIPS'13. Lake Tahoe: MIT Press, 2013: 3111-3119.