# The Distribution of Words in Chinese and Laos Based on Cross Language Corpus

Yuquan Huang[1,a], Lanjiang Zhou[2,b], Feng Zhou[2,c] and Jinpeng Zhang[2,d]

[1]School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
[2]Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming 650500, China

Email: [a]120166881@qq.com, [b]915090822@qq.com, [c]zf158@sina.com, [d]939127870@qq.com

**Abstract.** Word representation is the basic research content of natural language processing .At present, distributed representation of monolingual words has shown satisfactory application effect in some Neural Probabilistic Language (NPL) research, while as for distributed representation of cross-lingual words, there is little research both at home and abroad. Aiming at this problem given distribution similarity of nouns and verbs in these two languages, we embed mutual translated words, synonyms, super-ordinates into Chinese corpus by the weakly supervised learning extension approach and other methods, thus Laos word distribution in cross-lingual environment of Chinese and Laos is learned. We applied the distributed representation of the cross-lingual words learned before to compute similarities of bilingual texts and classify the mixed text corpus of Chinese and Laos, Experimental results show that the proposal has a satisfactory effect on the two tasks.

**Keywords.** Weakly supervised learning extension, cross-lingual corpus, cross-lingual words distribution representations, neural probabilistic language model.

## 1 Introduction

Lexical representation is an important part of Natural Language Processing. [3] It is an important technical support for information retrieval, data mining, knowledge mapping and other research directions. The goal of vocabulary representation method based on statistical machine learning [6] is to learn the probability representation of word sequences from natural language texts. In the process of training, each word sequence is different from the word sequences in the training corpus. In the process of monolingual glossary space representation of a traditional but effective method is N meta grammar model , it through the learning of target words a short window information to predict the probability of the target vocabulary. The disadvantage of it is that it can not reflect the influence of the word's words on the sequence generation probability and the

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

similarity of the distribution probability of the similar words. Neural probabilistic language model in monolingual environment to solve this problem well. Neural probabilistic language model by acquiring syntactic and semantic information of learning words represents the distribution of features from natural language text, the word sequence similarity is similar to the word distribution, validation of the word distribution can be well applied in POS tagging, named entity recognition, semantic role labeling and other natural language problems. Although the monolingual lexical distribution that has achieved good results, but in the cross language Natural Language Processing field research at home and abroad are scarce. There are two main methods at present: The first is the transfer of learning, the method will mark the learning information from one language to another language, so that fewer resources to obtain better processing results. The second approach is to translate the two languages into one or third languages, using a language to express cross language information.

The above methods have achieved certain results in dealing with cross language natural processing, but there are some disadvantages that cannot be transplanted, the algorithm process is complex, and the accuracy needs to be promoted. At present, the mainstream of the textual level analysis method only inspects the distribution characteristics of nouns and verbs. From this idea, in order to solve the above problems analysis of the distribution of Chinese, Lao noun, verb similarity, Lao nouns and verbs as nouns, verbs, Lao words into Chinese corpus, generating Chinese and Lao cross language lexical sequence corpus. Through the neural probabilistic language model to learn Laos nouns, verbs in cross language distribution in space. In this way in the cross linguistic data in learning the Chinese cross language word vector distribution, applied directly to the Laos text, learning resources and lack of resolve Laos corpus analysis of cross language text. The Chinese and Lao cross language text classification and text similarity based on the experiment, verify the Chinese and Lao cross language lexical distribution represents the effect.

## 2 Neural probabilistic language model

Neural probabilistic language model by Bengio Y[1,2] in 2003 first proposed joint probability function based on artificial neural network to learn the vocabulary of a language sequence. It solves the problem of sparse data dimensional and dictionary vector language of space effectively, and solve the distribution of N meta grammar model cannot solve the similar problems, and compared with the dictionary vectors and N meta grammar model can better represent the lexical distribution

Neural probabilistic language models are described as follows:

By a given word sequence: $\omega_1, \ldots, \omega_{t,}\ \omega_t \in \mathrm{V}$,V represents all the lexical sets of the target language. V although very large but finite, the goal of the neural language model is to learn a good function to estimate the conditional probability of a word:

$$f(\omega_t, \omega_{t-1}, \ldots, \omega_{t-n+2}, \omega_{t-n+1}) = p(\omega_t \left| \omega_1^{t-1}\right.) = p(\omega_t \left| \omega_{t-1}, \omega_{t-2}, \ldots, \omega_1\right.) \quad (1)$$

Among them, $\omega_t$ is the t word of the word sequence; The corresponding sub sequence written in $\omega_i^j = \omega_i, \omega_{i+1}, \ldots, \omega_{j-1}, \omega_j$ ; V represent for the thesaurus $|\mathrm{V}|$ represent the size of the thesaurus. The joint probability of word sequences can be obtained by conditional probability.

Formula (1) includes two processes: Firstly, the mapping $C$ is constructed to map any element in the set $V$ to the feature vector of the word $C(i) \in R^d$ .The distribution of feature vector which represents the word association Thesaurus . $d$ represents the dimension of feature vector. The free parameter matrix of $|V| * d$ is expressed in the experiment.

Then construct the probability function of word. We use the word feature vector C representation :Input context feature vector based on Corpus $(C(w_{t-n+1}) \ldots C(w_{t-1}))$ to prediction of probability distribution of the next word .The output of the $g$ is a vector ,The $i$ element of the vector is the estimated probability $\hat{P}(w_t = i \left| w_1^{t-1}\right.)$,Calculated by the following way :

$$f(i, w_{t-1,}\ldots, w_{t-n+2}, w_{t-n+1}) = g(i, C(w_{t-1})\ldots C(w_{t-n+2})C(w_{t-n+1})) \quad (2)$$

$f$ from the above mapping $C$ and $g$ combination, These two maps are related to some parameters. The parameter of mapping $C$ is the feature vector itself, which is represented as a $|V| * d$ matrix $C$ . $C$ line $i$ is the feature vector of the word $i$ . Function g can be realized by feedforward neural network or convolutional neural network. The formula (2) shows that the function of $f$ through the context words to predict the word I word eventually transformed into a distribution function of $g$ through the distribution of feature vectors to predict the $i$ context of a word.

## 3 The distribution of linguistic words in Chinese and lao languages

**The characteristics of the distribution of words in Chinese and Lao languages:** Chinese and Lao language have a large degree of similarity, they have a lot in common with the syntax. For example, in the same sentence: The syntactic structure of Chinese is (+attribute)subject + (+adverbial )Predicate +(+attribute)object(+complement);The syntactic structure of Lao language is (time/adverbial )+subject(attribute)+Predicate. The main two kinds of sentences, SVO sequence is completely consistent .The main difference is reflected in the Lao adverbial must be placed in the center after the word, the Chinese attributive adverbial, must be placed in the center of the word before. In terms of the composition of the sentence, the main content of the sentence reflects the main content. Attributive, adverbial, complement is not essential components of leaves. The two main components of SVO language is completely consistent. Subject, predicate, object, corresponding to the noun and verb in part of speech, the main structure of the sentence is the same. The distribution of the word sequence of the two language nouns and verbs should be similar.

Our goal is to ignore the differences between Chinese and Lao language，Let Lao language nouns and verbs as Chinese nouns, verbs, learning their distribution in the Chinese language environment, so that the mature Chinese text analysis method can be directly applied in Lao language text.

**Parallel corpus preprocessing:** We have selected 1026 pairs of parallel sentences which are obtained from China Radio Station and have been manually corrected. Although the original text contains all the text information, but the current Natural Language Processing [3] technology cannot fully handle these text messages. [3] Therefore, it is necessary to preprocess the text. Traditional text preprocessing is mainly to remove the stop words. Since this method requires sequence distribution on word learning, so we did not remove stop words, but we will have nothing to do with Chinese and Lao language text symbols, meaningless digital removal, and some of the names into a unified symbol, avoid because of names caused by the change of word sequence distribution of learning effect, reduce noise interference.

**Parallel corpus word alignment:** We mainly use cross heuristic algorithm, by running from Chinese to Lao language and Lao language mapping from the map to the Chinese two directions to get to the word alignment. [4,5]

We only consider the alignment of words in two directions. By word alignment, we can obtain a corresponding cross language [5] translation word in the parallel corpus.

Example sentence:

(1)We are going to/0 play basketball/1 this afternoon/2

(2)ตอนบ่ายนี้/2 ພວກເຮົາ/0 ຫຼິ້ນບ້ອງ/1

In the above of parallel sentences, the same suffix label of Chinese and Lao words indicate that they are aligned translation relations. Through GIZA$^{++}$ we can get the corresponding Chinese words in Lao words in Chinese instance. In Chinese examples, the Chinese word "we" will replace the ພວກເຮົາ generation for example as follows:

(3)ພວກເຮົາ are going to play basketball this afternoon

We place an example (1) and (3) into a neural probabilistic language model, get the distribution information of ພວກເຮົາ words in Chinese language environment. In the end, we can make the distribution of ພວກເຮົາ and Chinese words "we" are similar, in accordance with the law of linguistics, that is, the distribution of two words should be similar. We call this case (3) in the case of Chinese examples proper

position is embedded into the corresponding Lao word for the derived instance. The distribution of the cross language vocabulary can be obtained by learning the Chinese language examples and the derivative examples in parallel sentences together as the learning materials.

**Replace the Lao words and Chinese word similarity relation:** Generally if between two words is the dictionary translation relations will assume they are justified, but in the more general case, play similar syntax semantic roles of the word have similar word sequence of joint probability distribution. Although Hot and cold are a antonym, but in many contexts in the sequence distribution is very similar. In the case of today's weather hot and today's weather cold, there is a relationship: $P(hot|today, weather) \approx P(cold|today, weather)$.

**The massive Chinese and Lao language mixed corpus of weakly supervised learning process:**We take the Chinese examples and examples of mixed learning derived instances can get a certain degree of Lao Chinese vocabulary. But by the parallel sentences required for manual correction in a larger extent, limited corpus, learning the Lao lexical distribution still can not fully reflect the distribution of it should be in the Chinese language environment. For example, we can learn to ພວກເຮົາ with the word "we" similar relations in the mixed case, but the ພວກເຮົາ and the "we" in Chinese, "I", "we" should be similar, and they also have some similarities such as pronouns. The neural probabilistic language model can obtain the best results only in the large scale corpus. As a result, we have added the above mixed case set to the search of the Chinese news corpus, which has been collected in a large scale. We first concentrated in the large-scale corpus of Lao nouns and verbs corresponding Chinese nouns and verbs are replaced. Then replace the generation of derivatives, while retaining the original corpus of Chinese examples. Such as Lao "ພວກເຮົາ" in addition to "we" can be replaced, you can also replace "I", "we". Through this process we get the first mixed data set. We learn synonyms through cross language learning corpus for learning, can obtain Lao nouns, verbs in the Chinese language environment more accurate sequence distribution of information. Through this process "ພວກເຮົາ" and "we", "I" have made a similar distribution. In the neural probabilistic language model, if there is a similar context semantic or syntactic structure, there can be a similar distribution.

We are on the Lao word corpus contains a weakly supervised learning extension:

(1) The similarity of Lao words and Chinese words are compared, if the similarity is greater than the threshold value, we put Chinese words into candidate replacement word set.

(2) The Lao word and word candidate replacement word set into English, query semantic relations between them in the hierarchy tree English, if they are similar words or directly on the meaning of a word can be directly replaced, generate new candidate derived instances

(3) Whether the candidate derived from the artificial judgment is in accordance with the knowledge logic, and the derivation is reasonable, otherwise it will remove the case.

(4) The selected examples of corpus derived added concentration, through the neural probabilistic models of language learning new vocabulary in Chinese and Lao language distribution, and jump to 1.

(5) Repeat the process from (1) to (4) until you can't learn a new Chinese word.

Through the propagation process of weakly supervised learning vocabulary learning focused on the corpus, we found that with the increase of the size of the corpus, the similarity between Chinese words and vocabulary will decline slightly, but the similarity between Lao vocabulary and Chinese vocabulary increased significantly, and the expected results of similarity of the synonymous relations between Lao vocabulary and Chinese vocabulary is closer, and close to the similarity between words in Chinese.

(6) Learning model

Back propagation algorithm for learning model parameters in neural probabilistic language model. [2] At present, there are many parameter improved learning algorithm for the back propagation algorithm. We choose the improved gradient descent algorithm to optimize the parameters of the model set. The method can dynamically adapt to the first order information, and the computation cost of the gradient descent is minimal.

## 4 Experiment and analysis

**Text similarity computing method:** We first used neural probabilistic language [2] on cross linguistic corpus study, get the cross linguistic distribution of Chinese and Lao lexical representation. Based on the experience, we set the vector dimension of each word to 200,The number of neural units in the hidden layer

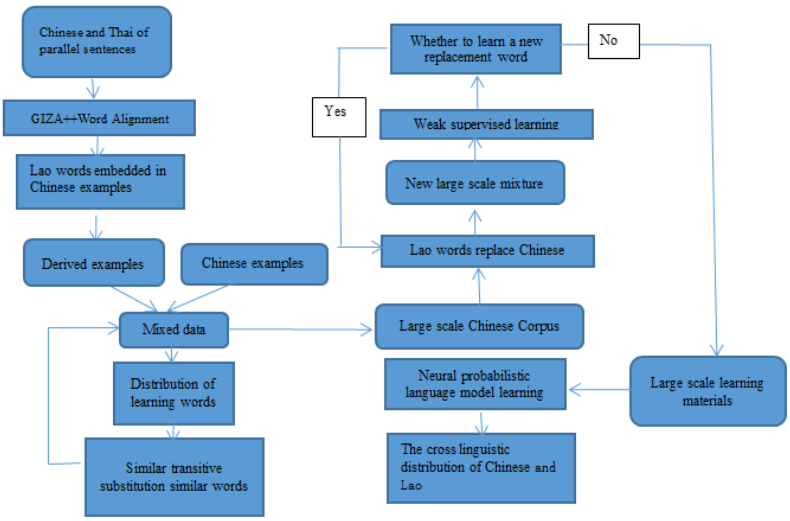
of neural probabilistic language model is 64 .The promise error is 0.001, and the training window is 5. The frequency is greater than or equal to 3 times the Chinese and Lao words only consider the process in the learning corpus. We have got a vocabulary learning based cross language distribution as text similarity computing. We have been the basis of learning Chinese and Lao language vocabulary distribution as text similarity computing.

We through the algorithm selected each document feature weight accounted for 5 of the feature words. The characteristic phrases for text t are $(\upsilon_{t1}, \upsilon_{t2}, \ldots, \upsilon_{t\xi})$, Weight as: $(\omega_{t1}, \omega_{t2}, \ldots, \omega_{t\xi})$, The features of the same text $k$ are $(\upsilon_{k1}, \upsilon_{k2}, \ldots, \upsilon_{k\xi})$,The corresponding weights of characteristic words are $(\omega_{k1}, \omega_{k2}, \ldots, \omega_{k\xi})$.The similarity between the two texts through the text t each feature word and text K in each feature word of the word vector cosine similarity and the respective characteristics of the weight of the multiplication and accumulation, and then divided by the total number of 25. The cosine similarity between the words $\upsilon_{k1}$ and $\upsilon_{t1}$ is expressed as $\upsilon_{k1}$ & $\upsilon_{t1}$. The formula for calculating the similarity of text:

$$t \ \& \ k = \frac{\sum_{i=1, j=1}^{\xi} \upsilon_{ti} \ \& \ \upsilon_{ki} * \omega_{ti} * \omega_{ki}}{2\xi} \tag{3}$$

**Experimental results and analysis:** We use Wikipedia on the Chinese and Thai text align text as the experimental text set, select the economic, political and technical sports, five kinds of Chinese and Thai parallel texts of 100 articles. The experiment consists of two parts. The first part: the Chinese and Lao parallel text similarity calculation; the second part: the text of Chinese and Lao mixed text concentrated randomly disrupted the order after the judgment in their classification in five categories. The similarity of Chinese and Lao text description of the cross language word distribution similarity between the two synonyms. Only the synonyms in the two texts are similar in the same vector space distribution, which can improve the similarity of the text.

The flow chart is as follows:




**The experimental results show that:** in the case of the same corpus size, the cross language word distribution in cross language text classification is better than two kinds of Machine Translation ways. For the expectation maximization algorithm considering the class information of the source language word translation as the biggest target language word translation probability, compared to cross language lexical similarity is the average similarity of all categories, higher accuracy, and the feature words combined with

semi supervised adaptation can update the target language text classification, the best effect. Experiments show that the accuracy of a cross language lexical distribution, lexical meaning expression accuracy. Based on the cross language word distribution, the classification knowledge of the source language can be directly migrated to the target language, which has a certain effect. [1,2]

## 5 Conclusion

In this work, To solve the cross linguistic distribution of Chinese and Lao words, we ignoring the differences between the two languages. We will place Lao nouns and verbs into Chinese corpus to generate cross linguistic data. Extension of corpus by weakly supervised learning, and then through the neural probabilistic models of language learning Chinese vocabulary representation of cross linguistic distribution. [2,3] The application of mature text analysis method can be applied directly to the Lao text in Chinese, and the application method in the analysis of cross language text on the elimination process is relatively simple, not divergent very complex. Experiments have achieved certain results through text similarity and text categorization. The next step is to improve the neural probabilistic language model to improve the accuracy of cross language lexical distribution. And further study the influence of the distribution feature vector representation dimension of cross language vocabulary on the representation of cross language lexical distribution.

## Acknowledgements

This paper is supported by National Nature Science foundation (No.61662040,61562049).

## References

1. Bengio S,Bengio Y. Taking on the curse of dimensionality in joint distributions using neural networks. IEEE Transactions on Neural Networks,2000,11(3):550-557.
2. Bengio Y, Ducharme R，Vincent P, et al. A neural probabilistic language model.Journal of Machine Learning Research,2003,4(3):1137-1155
3. Collobert R,Weston J,Bottou L,et al.Natural language processing(almost)from scratch.Journal of Machine Learning Research,2011,12(1):2493-2537.
4. Zeman D,Resnik P. Cross-language parser adaptation between related languages//IJCNLP,2008:35-42.
5. Sogaard A. Data point selection for cross-language adaptation of dependency parsers//Proc of the 49th Annual Meeting of the Association for Computational linguistics: Human Language Technologies:Short Papers-Volume 2,2011:682-686.
6. Ando R K,Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data.Journal of Machine Learning Research,2005,(6):1817-1853.