Chinese-Lao Bilingual Named Entity Alignment Research

Rui Han1,2, Lanjiang Zhou1,2, Feng Zhou1,2 and Jinpeng Zhang1,2

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

Corresponding author: 915090822@qq.com.

Abstract. Chinese-Lao bilingual NE alignment has a very important significance. Three entity alignment methods are proposed in this paper. Firstly, the paper proposes the similarity of bilingual entity fuzzy matching problem. Secondly, we use bilingual entity word sequence pattern similarity to propose Chinese entity model to match Lao entity method. Then we build a naive Bayes bilingual NE alignment model to align Chinese and Lao named entity in the comparable corpus, by mining knowledge information words of Chinese entities. In the end, the rules combine the advantages of the three methods are proposed to achieve the best results.

Keywords. Chinese, Lao, NE alignment, similarity, pattern matching, Naïve Bayes.

1 Introduction

Named entity refers to the entities which has a certain significance in the text [1], it is an important information in Natural Language Processing and an important support for information retrieval, automatic question answering, Machine Translation, topic discovery, and other information processing research [2]. For bilingual sequences, the alignment of bilingual entities aims at establishing the correspondence between the source language and the named entity in the target language and is an important task in the field of multi-language information processing such as Machine Translation, cross language information retrieval, etc. [3].

In recent years, the economic, political and cultural exchanges between China and Lao is deepening, which promote the development of Lao information processing technology, but also put forward higher requirements. But the research is still very weak in Lao languages.

Lao and Chinese Named Entities are quite similar, such as they both do not have special features like capitalization to help identify named entities. Moreover, in the sentence, there are not spaces to delimit the word. And the order of Subject, predicate and object is the same. Of course, it has its own characteristics. For example, if the personal name is the Lao local name, the first name is in front, the last name is in back, otherwise, the last name is in front, the first name is in back. In the sentence of Lao language, the adverbial is generally at last. The front of general location name of Lao has the special word to be distinguished. The personal name is represented as man if the front of name have Mr., the person name is represented as woman if the front of name have Mrs., etc. [4-5].

2 Selection of candidate equivalent entity

Candidate equivalent entity of Chinese named entity mainly were unknown words and identified similar entities in Chinese text in Lao parallel text. If the Lao parallel text is rare, consider comparable text with similarity more than 0.4. In the process of selecting candidate equivalent entities, we select the corpus first to find out the potential text set of
candidate entities, then we consider the similar entities and unknown words recognized from the text set as a set of candidate equivalent entities.

![Flow chart of screening candidate equivalent entities](image)

**3 NE alignment based on similarity of bilingual entity fuzzy matching**

To compare the similarity between Chinese named entity and the candidate entity selected by machine translation way, and translate the Lao candidate entity into Chinese by machine translation and compare the similarity between it and Chinese source entity. The higher similarity result between the two translation processes is the final similarity. If the similarity is greater than a certain threshold, we believe that the two are equivalent entities. Candidate entities are selected by making rules combined with unknown words. At present, Machine Translation has a poor translation of named entity, but some translation results can be used as reference. The correct part of the translation results can be referred to the entity screening and the similarity can be obtained by calculating the number of syllables and words accord with the feature of transliteration and paraphrase. The Lao candidate entity with the highest similarity and more than a certain threshold is the Lao equivalent entity.

The Chinese and Lao equivalent entities are determined by calculating their pronunciation consistency ratio, the specific calculation formula is as follows:

$$\text{similarity}(\text{Ech}|\text{Ela}) = \frac{\sum_{i=0}^{n} \text{Count}(\text{syichi}=\text{syila})}{\min(m,n)}$$

**4 NE pattern recognition**

Because of the differences between the Chinese and Lao pronunciation rules, the transliteration features are not fully applicable. Some Lao syllables have no corresponding pronunciation in Chinese which brings great obstacles to the named entity alignment based on transliteration.

Some entities do not fit in with the features of transliteration and paraphrase. For example, some people names often appear in the English, digital, abbreviated form such as Vill Wannarot. This paper argues that although there are differences between Chinese and Lao but in parallel corpora and comparable between Chinese and Lao entities in the corpus with the entity model should be the same or similar. We can get a high quality pattern that can meet the needs of people, place names matching by compiling statistics high frequency instance pattern through some common names,
translating the pattern into Lao and adjusting the order according to Lao language habits. Lao words which fit in with the pattern are equivalent entities. For example, Huang Hua, served as China’s foreign minister from 1976 to 1982. Translate the sentence into Lao and put the adverbial of time in the end of the sentence according to Lao language habits, generate a pattern. Compare this pattern with Lao text words sequence in the comparable corpus, and entity which fit in with the pattern is the equivalent Lao entity.

We can get a lot of Lao pattern by the manual adjustment pattern in Chinese. Lao entities mining through pattern have high accuracy in type judgement. For example, the most commonly used short pattern: [Mr. X says], the accuracy of matching Lao names in the corpus is 100%. If there is a need to align with Chinese entities, the pattern must contain sufficient context information. An equivalent pattern is shown as follows:

Chinese: Headmaster Yellen attends the meeting
Generate the following pattern:
Mr. X headmaster attend meeting
Headmaster can be replaced by similar name inspired feature words such as manager, doctor, etc.

5 NE alignment based on naïve bayes model

The work of the last chapter can achieve the alignment of entities in the simple sentences, but the matching effect of entities in some complex sentences is decreased. Therefore, this paper proposes statistical learning about the relevant knowledge information of the candidate entities, and Bayesian alignment through the knowledge information contained in the candidate words.

5.1 Knowledge information of names of people and place

A Chinese-Lao example sentence can be composed of a number of words. Consider words as data items and the assemblage of words is an item set. The occurrence frequency of an item set is the number of records that are included in the entire corpus, which is also called the degree of support of the item set. If the occurrence frequency of an item set is more than a certain frequency threshold value, it is the frequent item set. The set of all frequent set is denoted as \( L \); its frequency is \( \text{support} = \frac{\text{count}(L_k)}{N} \). The degree of support of this paper is set to 3.

Based on the study of this section, we recognize the Chinese named entities in Chinese-Lao corpus in a certain period. In the words sequence, nouns and entities that appears in a window of 8 words are knowledge information words related to the entity. Collect information words of the top three Co-occurrence degree with the entities and words whose degree of support is greater than 3, and form the two order frequent sets with entities.

5.2 NE alignment model based on Naïve Bayesian

Bayesian model is a kind of machine learning model, which is commonly used in the field of text classification and word sense disambiguation, and so on. It uses the prior probability of the category and the distribution of words to calculate the probability of unknown text or word meaning to a certain class. According to the Bayesian model, the correct equivalence entities should be the largest in the probability of the occurrence of the given knowledge information words. We define the decision rule \( S \) as follows:

\[
\text{if } P(E_k|C_{inf}) > P(E_{lao}|C_{inf}), E_k \neq e \text{ then decide e}
\]

Cinf is the knowledge information word sequence, Ek is any candidate Lao equivalent entity variables of a Chinese entity, Elao is a right Lao equivalent entity and \( P \) is probability.

The \( P(E_k|C_{inf}) \) in formula is unknown and needs to be calculated by the following formula:

\[
P(E_k|C_{inf}) = \frac{P(C_{inf}|E_k)P(E_k)}{P(C_{inf})}
\]
P(Ek) is the prior probability of candidate entity Ek, \( \frac{P(Cinf|Ek)}{P(Cinf)} \) is a correction factor in the knowledge information sequence, and the decision problem \( \frac{1}{P(Cinf)} \) is a constant which does not affect the final decision result.

In order to calculate the conditional probability \( P(Cinf|Ek) \) need to use Naïve Bayesian Assumption. Naïve Bayesian Assumption assumes that each knowledge information words are independent. So \( P(Cinf|Ek) \) can be calculated by the following formula:

\[
\text{Elao} = \text{argmax}_{E_k} \left[ \log P(Cinf|Ek) + \log P(Ek) \right] \tag{3}
\]

\[
P(Cinf|Ek) = P([K_i|K_i \text{ in } Cinf]|Ek) = \prod_{K_i \text{ in } Cinf} P(K_i|Ek) \tag{4}
\]

\( K_i \) is the information word in the sequence of knowledge information. Knowledge information words are independent with each other. The re definition of decision rules is as follows:

\[
\text{if } E_{\text{ao}} = \text{argmax}_{E_k} \left[ \log \sum_{K_i \text{ in } \text{knowledge}} P(K_i|Ek) + \log P(Ek) \right] \tag{5}
\]

\( P(K_i|Ek) \) and \( P(Ek) \) can be obtained from a Lao corpus using maximum likelihood estimation.

\[
P(K_i|Ek) = \frac{c(ek|k_i)}{c(ek)} \tag{6}
\]

\[
P(Ek) = \frac{1}{\kappa} \tag{7}
\]

6 Rules combining three methods

There are advantages and disadvantages of these three ways of entity alignment. The similarity matching method is the fastest and there is a certain percentage of the correct rate but the matching rate of the entity is very low. The pattern matching method’s accuracy is higher, but the recall rate depends on the quality of the existing model, the scale and the comparability of the comparable corpus. Bias alignment method has higher time complexity but can solve the equivalent entity that cannot be excavated by the former two methods. From the time complexity, the similarity fuzzy matching method is the lowest, the pattern matching method is higher, and the naive Bayes method is the highest. From the correct rate, the pattern matching method is the highest, and the similarity fuzzy matching is the lowest. So we use the following decision rules R2 mining equivalent entities:

Step 1. If the similarity between Chinese entity and the candidate Lao language entity exceeds the threshold value, then it is equivalent entity Ela1, otherwise the candidate result is null. And go to step 2.

Step 2. If the words sequence pattern matching between Chinese entity and the candidate Lao language entity is consistent, then it is equivalent entity Ela2, otherwise the candidate result is null. And go to step 3.

Step 3. Select the candidate entity with the maximum probability of the occurrence of the equivalent entity through the Bayesian formula and get the Lao equivalent entity Ela3, otherwise the candidate result is null. And go to step 4.

Step 4. Compare Ela1, Ela2 and Ela3. If the three are the same, we get the final equivalent entity Ela. If two of the three are the same, then take the entity same with the two. If none of them are the same, then take Ela2 as the final equivalent entity.

R2 combines the advantages of three equivalent entity mining alignment methods, and can dig out the Lao equivalent entities as much as possible and get the highest accurate rate.

7 Experiment and result
This paper sets up the bilingual corpus with Chinese-Lao parallel corpus and comparable corpus crawling from Internet. We have 214 Lao named entities and 214 Chinese entities corresponding to the Lao entities. In this paper, the effects of the three alignment methods are compared with the rule R2. The fuzzy similarity matching threshold is set to 0.4.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>Fscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity matching</td>
<td>62.17%</td>
<td>35.42%</td>
<td>48.34%</td>
</tr>
<tr>
<td>Pattern matching</td>
<td>71.69%</td>
<td>21.84%</td>
<td>34.46%</td>
</tr>
<tr>
<td>Naïve Bayesian</td>
<td>65.08%</td>
<td>67.69%</td>
<td>68.24%</td>
</tr>
<tr>
<td>R2</td>
<td>70.61%</td>
<td>73.93%</td>
<td>72.75%</td>
</tr>
</tbody>
</table>

The fuzzy similarity matching method can match equivalence entities of obvious characteristic of transliteration. Pattern matching method can match entities that does not have obvious characteristics of transliteration but the pattern sequence is close. Naïve Bayesian method can match entities that cannot be found by the first two methods, and mine the largest equivalent entities to achieve the highest recall rate. The combination of the three methods can get the most accurate entity alignment results.

8 Conclusions

Traditional word alignment approaches cannot come up with satisfactory results for named entity alignment. In this paper, we study Chinese-Lao entity common characteristic including translation similarity, pattern similarity and identical knowledge information. We propose three alignment methods based on similarity matching, pattern matching and Naïve Bayesian. Finally, the multiple matching rules are proposed by compromising the advantages of the three methods. Experiment results show the effect of the three methods each and the combination of the three methods and the method proposed in this paper obtained the ideal result under the present condition. The next step is to further improve the accuracy of bilingual NE alignment by making full use of more bilingual language features and some machine learning algorithm.

Acknowledgements

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

References