Translate Chinese Organization Names Using Examples and Web

Feiliang REN
Northeastern University
Shenyang, China
renfeiliang@ise.neu.edu.cn

Jingbo ZHU
Northeastern University
Shenyang, China
zhujingbo@mail.neu.edu.cn

Huizhen WANG
Northeastern University
Shenyang, China
wanghuizhen@ise.neu.edu.cn

Abstract:
This paper proposes a new approach for translating Chinese organization names that uses example-based method along with web assistance. It consists of two phases, first, it generates a translation candidate for the input Chinese organization name by an example-based translation method; and secondly, it uses the web to amend this translation candidate so as to finish such tasks: translation candidate reordering, word selection revising, and adjustment of the use of function words. Experimental results show that our method outperforms competing traditional statistical translation method in the task of translating Chinese ONs.

Keywords:
Chinese organization names translation; named entity translation; examples-based machine translation method; web assistant translation method; machine translation

1. Introduction

There are three main types of named entities (NEs): location names (LOC), person names (PER), and organization names (ONs). The translation of these types’ named entities is crucial for many NLP tasks, such as cross-language information retrieval, machine translation, and question and answering system. Among different kinds of NEs for Chinese, the ON’s translation is the most difficult part. And there are two main challenges in Chinese organization names’ translation. First, during translation, it is difficult to deal with Chinese organization name’s flexible structure, flexible translation order, and the problem of one word corresponds to multi translations even in the same kind of Chinese organization name. Secondly, many kinds of common Chinese organization names are less frequently occurred, they are more appropriate to be translated by traditional translation technologies than some search and extraction methods, but because of the expensive and time-consuming for constructing a large bilingual Chinese organization name corpus, the traditional statistical based translation method is hard to obtain reliable statistical parameters for translation, so it is difficult to obtain good translation results by these statistical based methods.

Looking at the previous methods on NE translation, we can classify them into three kinds: search and extraction based method, transliteration based method, and translation based method (such as word-based translation and phrase-based translation). Search and extraction based method usually needs to generate some queries first for the input text, and submit these queries to a search engine (such as Google or their in-house search engine), then extract the translation result from the returned documents (using some recognition and alignment technologies). This kind of method is appropriate for the translation of those named entities that often occurred on the web, such as the named entities that are among politics and news domain, or some famous companies’ name, or some important international events, and so on. But in real world, there are many kinds of less frequently occurring Chinese ONs (such as some local administrative departments, some small shops, and so on); it is difficult to generate some good queries for them to obtain their translations using this kind of method directly. For example, for such Chinese ON “抚顺市粮食局/Fushun Municipal Bureau of Grain”, its translation doesn’t appear on the web directly at all, so it is impossible to obtain good performance by this kind of method directly.

For transliteration method, it is very appropriate to translate foreign person name, and foreign location name. But among Chinese ON, most of these person name and location name can be either simply mapped as Chinese pinyin or a unique dictionary translation, the proportion of transliteration is very small. That is to say that in the translation of Chinese ON, the transliteration problem is not so significant.

For translation based method, such as statistical based translation method, it can obtain excellent translation results if there were a large bilingual training corpus. But for Chinese ON’s translation, it is very expensive and time-consuming to construct such a large training corpus. And without enough training data, it is impossible to obtain reliable statistical parameters for these statistical based translation methods. Besides, because of Chinese ON’s characters, the problems of reordering, word selection in the situation of one word corresponds to multi translation, and inserting or deleting
of some function words are still not discussed in these existed translation methods.

In response to these challenges, we present a two-step approach to translate Chinese ON. It uses example-based translation method to solve the problem of less large training corpus, and uses the web to solve the problems correspond to Chinese ON’s characters. In the first step of this method, it generates a translate candidate for an input by an example-based translation method, and then in the second step, it uses the web to amend these translation candidates so as to finish such tasks: reordering, word revising, and the inserting or deleting of some function words.

Note that in this paper, our focus is on translating Chinese ON out of context.

The paper is organized as follows, section 2 presents related work; section 3 describes our translation approach in detail; section 4 provides our experiments and finally section 5 concludes the paper.

2. Related Works

Up to now, there are a lot of excellent works have been done on the research of NE translations based on different approaches. Huang et al.[1] proposed an approach to translate rarely occurring NE based on search and extraction. For an input NE, their method first generates some queries in the target language according to some contextual information, and then searches for relevant documents from a target language corpus; and the final translation result is extracted from the selected documents based on phonetic and semantic similarities. Similar works can also been found in the works of NAGATA[2], Wu et al. [3], and Alegria et al.[4], which use the web as a large corpus and search and extract the translation results from web.

For the transliteration approach, Stalls and Knight [5] introduced a back transliteration model. And Al-Onaizan and Knight [6] proposed an algorithm based on sound and spelling mappings using finite state machines. Huang [7] proposed and cluster-specific transliteration method, which for an input NE, it first classified it into the most likely cluster, and then transliterate this NE with the corresponding models.

For the translation approach, IBM Model 1[8] is usually used during the translation, such as in Al-Onaizan and Knight’s work[6], and in Hassan and Sorensen’s work[9]. Besides, many phrase-based translation approaches have been proposed. Huang et al. [10] proposed an approach to translate NE based on the minimization of a linearly combined multi-feature cost. And similar work can also been found in Moore’s work [11].

The most similar prior work with our method is Al-Onaizan and Knight’s work[6], that first produced some translation candidates by transliteration model and a modified IBM Model 1, then used the web as a re-scoring tool to select a final translation result. But in their work, the web is mainly used as a re-scoring metric.

3. Our Two Step Approach

The whole translation process in our approach is carried out in two main steps. Given a Chinese ON, our approach first generates a translation candidate using an example-based method; and then, amends this candidate using the web so as to finish the tasks of reordering, word revising, and the inserting or deleting of some function words for the translation candidates.

3.1. Generating Translation Candidate by Example-Based Method

As described in the introduction, the translation of Chinese ON is the most difficult part among the three general NE types, this is not only because the inherent characters of Chinese ON, but also because the expensive and time-consuming for constructing a large bilingual training corpus. The lack of enough training corpus prohibits the use of some statistical based translation method. But we know for the example based translation method, as long as there was one translation example whose similarity with the input text is within an acceptable predefined threshold, the system can output a good result. This is the motivation for us to use an example-based method to produce the translation candidate.

Example-based method does translation based on the principle of analogy. Here in our system, we use an example-based translation method that is similar with the method proposed by Ren et al.[12]. In their method, the whole translation process is carried out in two main steps: matching step and generation step. In matching step, a most similar translation example with the input is selected for the next generation according to a cosine similarity metric. And then in the generation step, some operation states are extracted from the input text and the selected translation example, then an automaton is constructed to finish the final translation result’s generation.

We use an approach similar to that presented by Ren et al. But to make their method appropriate to our task, we modify their method from the following two aspects.

First, to obtain good translation performance, their method needs a word alignment result with high precision. This is possible when the training corpus is large enough. But for the task of translating Chinese ON, this condition is hard to be met. To overcome this problem, we take the English-Chinese bilingual LDC dictionary as part of the training corpus and add them in our limited Chinese ON corpus. In our system we train
the word alignment using GIZA++\(^2\) in both directions between source and target. And take the intersection of the two alignments as final alignment result. Besides, because of the problem of data sparseness, many sentence pairs would obtain NULL alignment results, in our system, we remove all these NILL alignment results, and take the rest as the final translation example corpus.

Secondly, in Ren et al.’s work, the input’s translation is generated only using one most similar translation example. This is feasible when the training corpus is large enough. But for the translation of Chinese ON, it is always that the training corpus is not large enough, so in many times, it is hard for the system to find a similar example that subject to the restriction of threshold for the input text. To overcome this problem, here we modified their original method and use multi translation example to generate the input text’s translation. To do this, we must segment the input text into several pieces when needed, translate these pieces respectively, and then recombine these pieces’ translation together to generate the final translation result. This requires that there must be at least one appropriate translation example in the corpus for every piece. With the restriction of this condition, we hope the segmented pieces’ number should be minimal.

In our method, system will fist try to find one appropriate translation example from the training corpus based on the method of Ren et al.’s, if failed, system will remove the last word of the input and take the remain string as a new input, and translates this new input according to original example-based translation method; if succeeded, system will take the rest untranslated string as new input text and repeat the previous process; if failed again, then remove this string’s last word again, and try to translate the modified string based on original translation method again. Repeat this process until all the input’s words have been translated. Finally, we need to recombine these translations into a unified text. According to the third character of Chinese ON, most of Chinese ON can be translated sequentially, we here simply join these translation results in their source language’s order. The following example demonstrates this translation process.

Suppose there is an input Chinese ON “调查 卢旺达 境内 人权 情况 特别 报告员 特别 报告员/ Special Reporter on the Situation of Human Rights in Rwanda” to be translated. System will first try to find an acceptable translation example for it (supposing the similarity threshold is set as 0), but failed; then system remove its last word and take “调查 卢旺达 境内 人权 情况 特别” as a new input string and repeat the matching process and try to find an acceptable translation example for it, but failed again; then system will remove the current string’s last word “特别” and take the rest string “调查 卢旺达 境内 人权 情况” as new input string and repeat the above process; until the input string becomes “调查 卢旺达 境内 人权”，system find an appropriate example “卢旺达 护卫 人权 联盟/ League for the Defense of Human Rights in Rwanda”, using this translation example, system will generate a translation “human rights for the research in Rwanda internal”; and then the system will take the rest untranslated string “情况 特别 报告员” as a new input string and repeat the above process again. And finally system will generate its translation “Special Reporter on the situation” based on the selected translation example “教育 权利 特别 报告员/ Special Reporter on the right to education”. The reader is referred to Ren et al. [12] for detail description of this translation generation process.

And finally, the input text’s translation is generated by simply joining these two fragments’ translation together in their source language’s order: human rights for the research in Rwanda internal Special Reporter on the situation.

This translation result is not so fluent. This is a specific problem called “boundary friction” in an example-based translation method (see [13],[14]), which means system generates a number of partial translations which are then pieced together rather than immediately producing one overall translation, this is robust to ill-formed input, but is subject to disfluency at phrasal translation boundaries even for well-formed input. The “boundary friction” problem is mainly caused by the wrong translation order and the wrong inserting or deleting of some function words. For example, for the above translation result, if we change its order and edit its function words, we can obtain such a translation “Special Reporter on the Situation of Human Rights in Rwanda Internal”, and this is a very good translation.

So we know that to obtain a good translation result using example-based translation method, we must solve the “boundary friction” problem properly.

3.2. Amend Translation Candidate Using the Web

Except for the “boundary friction” problem, there are other two problems unsolved in the translation of Chinese ON.

First, even the most robust translation method can hardly distinguish the situation of when to translate “东北” to “Northeastern” as in “东北大学/Northeastern University”, when to translate it to “Northeast” as in “东北林业大学/Northeast Forestry University”, and when to simplify map it as Chinese pinyin “Dongbei” as in “东北财经大学/Dongbei University of Finance and Economics”.\(^3\)

\(^2\)http://www.fjoch.com/GIZA++.html
Second, it is very difficult for a translation method to tackle the reordering problem and inserting function words problem in the translation of Chinese ON. Such as the translation of Chinese ON “中国建设银行/China Construction Bank”, “中国农业银行/Agricultural Bank of China”, and “中国建设银行/China Construction Bank”, and so on.

In fact, the “boundary friction” problem can also be viewed as the problem of reordering, and the problem of inserting or deleting of some function words.

Thus we can classify all the problems during translating Chinese ON into three types: the problem of reordering, the problem of inserting or deleting some function words, and the problem of word revising (we think in the situation of one word corresponds to several translation, system may make a wrong decision, but by word revising, we still can obtain the right translation.).

It is difficult to solve these problems directly, because there are almost no rules for them. But we know for a search engine (such as Google, Yahoo, Live, and so on), when we submit a query, it will return some exact matching items along with some partial matching items which are similar or relevant with the input query. If we take the obtained translation candidate as a query and submit it to a search engine, it is very possible that the true translation result will be included in the returned pages. For example, for a Chinese ON “中国农业银行”, the obtained translation candidate is “China Agricultural Bank”, when we submit it to Google, its true translation result “Agricultural Bank of China” is among the returned pages. This inspired us using the web to amend the obtained translation candidate so as to solve the above problems.

The whole amending process in our approach is carried out in three main steps: query, extraction, and refinement.

**Query** Take the obtained translation candidate as a query, and submit it to a search engine.

**Extraction** Exact some fragments that satisfy following condition1 or condition2, and satisfy condition3 at the same time.

Condition1: Extract a minimal fragment that satisfies:
- The substantive words are completely matched between this fragment and the query.
- The extracted fragment doesn’t contain any words that are different with the query except for some articles or prepositions.

Condition2: Extract a minimal fragment that satisfies:
- It has a same substantive words’ number with the query.
- There is at most one unmatched substantive word between it and the query. Except for this unmatched substantive word, it doesn’t contain any words that are different with the query except for some articles or prepositions.

For the unmatched substantive word, there must be at least one dictionary translation (omit the Chinese function characters, such as “的/de”, “了/le”, “着/ze” and so on) that appears in the original Chinese ON.

Condition3: For every query, we extract at most k fragments.

Among these conditions, condition1 is designed to solve the problems of reordering, and inserting or deleting of function words. Condition2 release condition1’s restriction, which aims to solve the problem of word revising for the situation like translating Chinese word “东北”. Here we only permit one word be revised by the web, because we think if there were too much words revised, the final translation performance is hard to be guaranteed. And condition3 is designed to guarantee the whole system’s translation speed.

**Refinement** Take these k fragments and the query as candidates, and submit them to search engine respectively, take the returned web count as a scoring metric, and the one that has the largest web count will be selected as final translation result.

The following two examples demonstrate the whole amending process (to simply the demonstration process, in following two examples, we set k in condition3 as 1).

**Example1**: Suppose an input Chinese ON “东北大学”, the obtained translation candidate by the example-based method is “Northeast University”, we use Google as search engine to amend this candidate.

First, submit “Northeast University” to Google;

Second, among the first returned page, according to extraction condition2, we can extract a string “Northeastern University” from the text “Welcome to Northeastern University, China”. There is one unmatched word “Northeastern” between the query and the extracted text, and we look up this unmatched word in a dictionary and find it has a translation “东北的”, after removing the function Chinese character “的”, the rest words “东北” can be found occurring in the original input text “东北大学”, so we take “Northeastern University” as an extracted item.

Third, we submit “Northeastern University” and “Northeast University” to Live respectively again, and their web counts are:

Northeast University: 27,700200
Northeastern University: 2,330,000

So we take “Northeastern University” as the final translation result. From this example we can see that using the web, we can solve the problem of word revising on the situation of one word corresponds to several translations.

---


---
Example2: Suppose an input Chinese ON “中国农业银行”, the obtained translation candidate by the example-based method is “China’s Agriculture Bank”, we use Google as search engine to amend this candidate. First, submit “China’s Agriculture Bank” to Google; Second, among the first returned page, according to extraction condition1, we can extract a string “Agriculture Bank of China”. The substantive words between this string and the query are completely the same. The different parts are some function words and the word’s sequence. Third, we submit “Agriculture Bank of China” and “China’s Agriculture Bank” to Google respectively again, and their web counts are:

Agriculture Bank of China: 27,100  
China’s Agriculture Bank: 86  
So we take “Agriculture Bank of China” as the final translation result. From this example we can see that using the web, we can solve the problem of reordering and inserting of function words.

From these two examples we can see that using the web to amend translation candidates is an effective method to solve the problems met during translating Chinese ON.

4. Experiments and Discussion

In the following comparison experiments, we evaluate our method from two aspects. First we evaluate the effectiveness of example-based translation method by comparing it with Moses5, which is a state-of-art statistical phrase translation tool. Second we evaluate the effectiveness of using web to amend the translation candidate. We use NIST score and BLUE value as evaluation metrics. And the evaluation tool is NIST MT Evaluation Toolkit6.

4.1. System Resources

Bilingual Corpus Totally we download 20729 bilingual Chinese ONs from the web7. From these bilingual Chinese ONs, we selected 300 sentence pairs randomly as test set, and take the Chinese part as input and the English part as their corresponding reference set. For the rest part, we take them as both the training corpus for Moses and the example corpus for our example-based translation system. Some statistics of these training corpus and test data are listed in table1. It is worth mentioning that using only one single reference to evaluate the system’s performance has limitations, many good translations are considered wrong because they do not exist in this single reference. However, we think the performance’s trend it reflected is still credible.

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td># of sentences</td>
<td>20429</td>
</tr>
<tr>
<td># of Chinese words</td>
<td>82887</td>
</tr>
<tr>
<td>Average # of Chinese per sentence</td>
<td>4.06</td>
</tr>
</tbody>
</table>

Bilingual Dictionary Bilingual dictionary is a necessary element for example-based translation method. Here we used the word-based C-E translation lexicon from LDC as bilingual dictionary for translation. Besides, this bilingual dictionary is also used as extra bilingual corpus to train word alignment as described in previous section 3.1.

Language Model Language model is needed for both Moses and Ren et al.’s translation method. Because the available bilingual Chinese ON corpus is very small, it is hard to train a robust language model using it, even if we added a bilingual dictionary into this corpus. And to make our work representative, we use Europarl Parallel Corpus8 to train a language model with the toolkit SRILM9.

4.2. Experimental Results

Our experiments are designed to evaluate our method from two aspects: the effectiveness of example-based translation method, and the effectiveness of using web to amend translation candidate. We take Moses as our baseline system, and experimental results are shown in table 2. In this table, “Example Method” means the example-based translation used in our system, and “Moses + Web” means using the web to amend Moses’s translation results, “Example + Web” means using the web to amend the translation result generated by example-based translation method. From table 1 we know, the average length of a Chinese ON is close to 3, so we using BLUE-3 as the evaluation metric.

<table>
<thead>
<tr>
<th>Translation</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses</td>
<td>0.1104</td>
</tr>
<tr>
<td>Example-based Translation Method</td>
<td>0.1970</td>
</tr>
<tr>
<td>Moses + Web</td>
<td>0.1761</td>
</tr>
<tr>
<td>Example + Web</td>
<td>0.2358</td>
</tr>
</tbody>
</table>

From table 2, we can see that example-based generated better translations compared with Moses in Chinese ON’s translation. This is mainly because there are no enough training data for Moses, so it is very difficult to obtain reliable statistical parameters. But for example-based translation method, as long as there is one acceptable translation example in the corpus, it can

5http://www.statmt.org/moses/  
6http://www.nist.gov/speech/tests/mt/resources/scoring.htm  
7http://www.fane.cn/dic/No3/300~358.htm  
8http://www.statmt.org/europarl/  
9http://www.speech.sri.com/projects/srilm/
use this example to generate a good translation result. We can use an example to demonstrate the difference between these two translation methods.

Suppose there is a Chinese ON “全国 捍卫 民主 理事会/National Council for the Defense of Democracy” to be translated. But in corpus, there is only one similar example with the input which is “全国 民主 阵线/National Democratic Front”, and it not only has some common substantive words with the input, but also has a similar length with the input. For the example-based method, this is enough to generate a good translation result, and Figure 1 demonstrates the translation process according to Ren et al.’s example-based translation approach[12]. But for a statistical translation method, only this one similar example is far from enough to train some necessary parameters, so it is impossible to obtain good translation result using these insufficient parameters. This can also be verified from Moses’s translation output for the input Chinese ON, which is “All Defense of Democracy Board”. It is obvious that this result is not so good as the example-based method’s output.

Figure 1. An Example of Translation by Example-based Method

Because of the data sparseness, there are a lot of such situation that no enough training data for the statistical translation method to train some reliable statistical parameters. This is the reason why statistical translation method can obtain best translation performance in many MT Evaluation Tasks, but obtain worse translation performance compared with the example-based translation method in the translation of Chinese ON.

From table 2, we can also see that when using the web to amend the translation results, both Moses and the example-based translation system make a great improvement on the translation performance. Many mistakes in the original system’s output have been revised by the web. Table 3 shows some translation examples using example-based method before and after using the web. For Moses, it obtains the similar results as shown in table 3.

From table 3 we can see that before using the web, there are some mistakes among these translation candidates compared with the reference, but after using the web to amend these candidates, these mistakes are corrected. This indicates that our method is effective for solving the problems met during the translation of Chinese ON. Of course, there are still many candidates that can not be revised rightly, such as for the Chinese ON “调查 卢旺达 境内 人权 情况 特别 报告员 /Special Reporter on the Situation of Human Rights in Rwanda”, its translation candidate is “human rights for the research in Rwanda internal Special Reporter on the situation”, because there are not any fragment that can be extracted based on extraction condition 1 or condition 2 described in section 3.2, we can not amend it.

Table 3: Some Examples of Using the Web to Amend the Translation Results

<table>
<thead>
<tr>
<th>Input</th>
<th>Ref</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>中国银行</td>
<td>Bank of China</td>
<td>China’s Bank</td>
<td>Bank of China</td>
</tr>
<tr>
<td>东北大学</td>
<td>Northeastern University</td>
<td>Northeast University</td>
<td>Northeastern University</td>
</tr>
<tr>
<td>东北财经大学</td>
<td>Dongbei University of Finance and Economics</td>
<td>Northeast Finance and Economic University</td>
<td>Dongbei University of Finance and Economics</td>
</tr>
</tbody>
</table>

We also notice that example-based translation method obtain more performance gain than Moses when using the web to amend the translation candidates. This is because that according to the extraction condition 1 and extraction condition 2, only the translation candidates that have at most one unmatched substantive word with its possible true translation can be revised by the web, these conditions are so strict that only the system that can generate good translation even by itself may obtain better translation result by the further result revising. That is to say, the better a system can generate a translation result, the more extra performance gain may be obtained by the web amendment method.

The experimental results also indicate that our method of using the web express very high ability of solving the problems of reordering, and inserting or deleting of some function words. For the problem of word revising, to avoid introducing some noise data on the web, our conditions are very strict, this guarantee that as long as it can revise a word, the reliability will be very high.

5. Conclusions and Future Work

This paper presented a Chinese ON translation method that used examples and the web. The whole translation process in this approach is carried out in two main steps. Given a Chinese ON, our approach first...
generates a translation candidate using an example-based method; and then, amends this candidate using the web so as to finish the tasks of reordering, word revising, and the inserting or deleting of some function words.

Currently, our conditions of using the web are very strict. For condition 1 and condition 2, as long as one of them is not satisfied, our method will not carry out any operation. But we know, even if these two condition are not satisfied, there are still some very useful information on the returned page for solving the problems of reordering, and inserting or deleting of some function words. In the future, we would take this information into account, and use them to improve the performance of reordering, and inserting or deleting of some function words.

Acknowledgements

This work was supported by the Open Project of National Laboratory of Pattern Recognition (NLPR), and was also supported in part by National Science Foundation of China (60873091), Natural Science Foundation of Liaoning Province (20072032) and Shenyang Science and Technology Plan (1081235-1-00).

References