An EBMT System Based on Word Alignment

HOU Hongxu, DENG Dan, ZOU Gang, YU Hongkui, LIU Yang, XIONG Deyi and LIU Qun

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, PR China
hxhou@ict.ac.cn

Abstract

This system is an experiment of examples based approach. It is based on a corpus containing 220 thousand sentence pairs with word alignment. The system contains four parts: matching and search, fragment matching, fragment assembling, evaluation and post processing. We use word alignment information to find and combine fragments.

1. Introduction

This system is our first experiment of example based approach. It is based on corpus with word alignment. The corpus contains 220 thousands of news, literal, dictionaries, dialog sentence pairs. All sentence pairs are POS tagged and word aligned.

2. System Architecture

The system has two parts: corpus and program. The corpus includes 220 thousand sentences pairs and a 460 thousand words and phrases dictionary. The program has four parts: matching and search, fragment matching, fragment assembling, evaluation and post processing.

2.1. Matching and searching

Search for the most similar sentences from corpus.

2.2. Fragment matching

Find out all matching and non-matching fragment of example sentence and corresponding fragment of translation of example sentence.

2.3. Fragment assembling

Assemble the fragment into a full sentence.

2.4. Evaluation

Evaluation the result of translation and determine keep or discard a non-aligned part.

2.5. Post processing

Process spaces, cases and punctuations.

3. Corpus

This system uses a corpus having 220 thousand sentences pair. It contains news, literal, dictionaries, dialog, etc.
The sentences in corpus are POS tagged and word aligned.

3.1. POS tag

The source language of the corpus is Chinese, and target language is English. For Chinese POS tagging, we use the ICTCLS 2.0. It's developed by ICT. It uses Multi Layer HMM. In former test of Chinese High Technology Development and Research, it got 97.58% of accuracy and could process 31.5Kb characters per second.

3.2. Word alignment

Word alignment is the base of our system. The arithmetic of word alignment is based on dictionary. It uses large scale bilingual dictionary, Word-Net and other human-readable dictionary. It is inspired by Ker's method [4]. This method mainly depends on similarity measured by bilingual dictionary, relative distortion information and Part-of-Speech information to align words. By setting alignment window it acquires many-to-many word alignments. On a test set of 650 translation sentence pairs of Chinese and English, in which Chinese sentence has 24.8 words in average and English 34.5, the word alignment system gets a result of recall 62.9 ％ at the precision of 84.0%.
Our algorithm is improved on Ker's in these aspects:
(1) The computation of relative distortion of Ker is improved, and the initial alignment anchors chosen by dictionary-based word similarity is added to improve alignment.
(2) Proposed a concept of 'alignment window'. By setting alignment window in the aligning process, many-to-many word alignments can be found.

3.3. Dictionary

The system uses a 460 thousand words and phrases bilingual dictionary. The dictionary is the base of word alignment and translation.

4. Matching and searching

This step is searching for the most similar sentence pairs from example base. There are two problems: how to measure the similarity of two sentences and how to find out the most similar sentence from a large scale corpus.

4.1. Measure of similarity

The measurement of similarity determines whether or not find out the most fitted example for translation. Because our system is based on word, we must find out the longest match fragment for translation. We use follow formula to measure the similarity,

$$m = \sum w(pos(i)) * match(i) * w2(i)$$

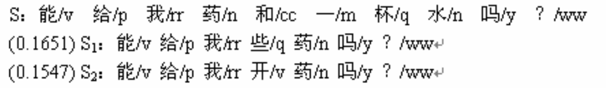
The first item *w* is the weight of POS, different POS has different weight. In our system, verbs have the largest weight and stop words and named entities have the lowest weight.

The second item *match* is 1 if the corresponding words are matched and 0 if not.

The last item *w2* is the measure of concatenation of words. If there are more concatenated words matched, *w2* value is larger. In our system the value of *w2* calculated as follow,

$$w2 = l^2$$

l is the length of matching string.



S：能/v 给/p 我/rr 药/n 和/cc 一/m 杯/q 水/n 吗/y ？/ww

(0.1651) S_1：能/v 给/p 我/rr 些/q 药/n 吗/y ？/ww↵

(0.1547) S_2：能/v 给/p 我/rr 开/v 药/n 吗/y ？/ww↵

Figure 1: An example of similarity.

Figure 1 give out a example of similarity, for source sentence *S*, there are two most similar sentences S_1 and S_2, and the similarity of S_1 is 0.1651 and the similarity of S_2 is 0.1547.

For the similarity calculation is bi-direction, the largest value is 2.0 and the lowest values is 0.

4.2. Searching

This step is the most time consuming step of the system. Actually, the efficiency of the step determines the efficiency of translation.

In our system, an index is created for very words appeared in corpus. So, we can find all sentences which contain certain words.

In the system, we search the example base for each word in source sentence orderly, and the searching results are joined into a set.

Through some experiments, we know some words are not helpful for finding out the most similar example and consume much long time. So, we exclude highest frequency words from searching.

After searching, the most similar examples are selected as final candidate set.

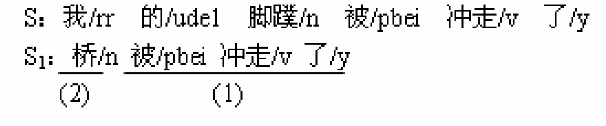
5. Fragment matching and assembling

The key of EBMT system is how to split sentence into fragments and how to assemble the fragments into sentence. The familiar methods are based on parsing or word. We adopt word-based method.

5.1. Fragment matching

This step is finding out all matching or non-matching fragments from source sentence and example sentence.

First, we find out all matched words, and then find all mismatched words but have same part of speech.



S：我/rr 的/ude1 脚蹼/n 被/pbei 冲走/v 了/y

S_1：桥/n 被/pbei 冲走/v 了/y

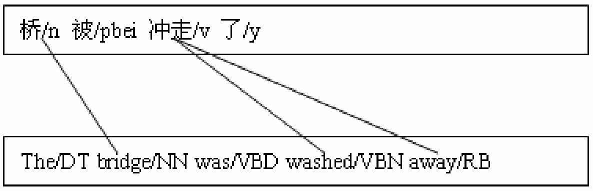
(2) (1)

Figure 2: An example of fragment matching.

Figure 2 shows a procedure of matching. *S* is the source sentence, and S_1 is the examples. First, the matching word string "was washed away" is matched (labeled as 1), then "bridge" (labeled as 2) is not matched, but there is a noun "web" is funded, so "bridge" is matched. In this example, there is only one matched fragment, because (1) and (2) are concatenated.

5.2. Target fragment

After finding out the matched fragments, we must find out their corresponding fragment in target example. The word alignment is the guild of this step.



桥/n 被/pbei 冲走/v 了/y

The/DT bridge/NN was/VBD washed/VBN away/RB

Figure 3: Word alignment.

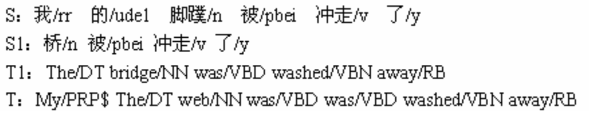
Figure 3 shows an example of word alignment. The upper sentence is source example, and lower sentence is corresponding target sentence. There are 3 pairs, ("qiao", bridge), ("chongzou", washed) and ("chongzou", away). As an example, we assume that "qiao" and "bei chongzou le" are different matched fragments. The first fragment "qiao" has one aligned pair, then "The bridge was" is the corresponding target fragment. The second fragment "bei chongzou le" has two aligned pairs, then "was washed away" is the corresponding target fragment. In this example, the word "was" is contained in both target fragments. It will be processed in later step.

5.3. Fragment assembling

After finding out the target fragments, there are several matched target fragments and non-matched target fragments.

For non-matched fragments, it must be searched in candidate set again for most similar examples. If can not find more example, it will be translated using dictionary.

Actually, word alignment determines the positions of target fragments.



S：我/rr 的/ude1 脚蹼/n 被/pbei 冲走/v 了/y

S1：桥/n 被/pbei 冲走/v 了/y

T1：The/DT bridge/NN was/VBD washed/VBN away/RB

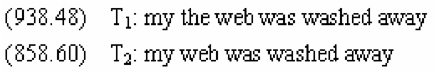
T：My/PRP$ The/DT web/NN was/VBD was/VBD washed/VBN away/RB

Figure 4: Fragment assembling.

Figure 4 shows an example of fragment assembling. Word matched fragment "was washed away" is placed in the target sentence directly, and the POS match fragment "bridge" is replaced by "web"--the translation of "jiaopu".

5.4. Make choices

In the last example, we could find the target sentence *T* contains many extra words, such as "The", "was". They may be the correct parts of the target sentence or wrong. So we must make choices. We use N-gram to determine which words will be hold and which words will be discarded.


(938.48) T_1: my the web was washed away

(858.60) T_2: my web was washed away

Figure 5: An example of N-gram.

Figure 5 shows an example of N-gram. The first sentence contains a wrong word "the", and in the second sentence it be removed. The perplexities of two sentences are 938.48 and 858.60, clearly, the second sentence is better than the first one. Then, the second sentence will be the last translation result.

6. Evaluation results

Because of carelessness in submitting translation result, the official result of our system is very bad. Later, we have submitted the result with correct format.

Table 1: The score of official and correct results

	Official result	Correct result
BLEU	0.0798	0.2013
GTM	0.3862	0.6380
NIST	3.6443	6.4716
WER	0.8466	0.6275
PER	0.7650	0.5187
Fluency	2.7180	
Adequacy	3.0820	

These two result are generated by same corpus and engine, but the correct result processed cases, punctuations etc.
We can divided the translation result into four classes,

(1) Existed in corpus
There are about 8% of test sentences have already existed in the corpus. These sentences are translated extremely well.

(2) Good fragments of a example
The test sentence is a part of example, and well aligned, so we can easily get the translation sentence from examples. The translations of these sentences are very good.

(3) Replace some words
These sentences are very similar to the examples, and because there are only a few words are different, the examples can be found out from corpus. The most errors of translations are the chosen of words.

(4) Part of example
Like (2), the sentence is the part of example, but the example is not well aligned or lost some important part. The most errors of translations are lost or add words.

(5) Combine of phrases
The sentence is combined by several phrases that could be found in different examples. The most errors of translations are lost or add words.

(6) Small fragments
Theses sentences have no similar examples in corpus, so they are often translated using dictionary. The translations of these sentence are often very bad.

7. Discussion

We have only completed a rudiment of an EBMT system. It has main elements of EBMT, but there are many things to be improved.

7.1. Phrase Recognition

In our system, a fragment splitting is totally blindfolded. So there are many half-balked fragments. To avoid this, we need phrase recognition to determine which positions may be divided and which position may not.
Because there are many phrases in the dictionary, phrase recognition should improve the efficient of phrase in the dictionary.

7.2. Word alignment

However, the accurate and recall of word alignment is still low, especially recall. So, there are large number of words need to be choose by N-gram. We must improve the arithmetic of word alignment.

7.3. Word cluster

Word cluster will be helpful for rising the accurate of sentence matching.

7.4. Sentence classify

Actually, divided examples into different class will reduce the time consuming of searching and rise the accurate of sentence matching.
We can divide examples into questions, negatives and so on.

7.5. Corpus

Our corpus is not complete checked for correction and fitness. There are many repeated or out of season contents.

8. Conclusions

After completed this experiment system, we must improve our system. An idea is combining EBMT and SMT. We have begun some attempt of SMT and other improvement.

9. References

[1] Satoshi Shirai, Francis Bond and Yamato Takahashi. "A Hybrid Rule and Example-based Method for Machine Translation".

[2] Lambros Cranias, Harris Papageorgiou, Stelios Piperidis. "A Matching Technique in Example-Based Machine Translation".

[3] Ying Zhang, Ralf Brown, Robert E. Frederking. "Adapting an Example-Based Translation System to Chinese".

[4] Sue J. Ker, and Jason S. Chang. "Align more words with high precision for small bilingual corpora"[J]. Computational Linguistics and Chinese Language Processing, 1997, 2(2):63-96