Auto Word Alignment Based Chinese-English EBMT*

Yang Muyun, Zhao Tiejun, Liu Haijie, Shi Xiaosheng and Jiang Hongfei

Research Center for Language Technology,
School of Computer Science and Technology,
Harbin Institute of Technology,
Harbin, China.
{ymy,tjzhao,hjliu,xshshi,hfjiang}@mtlab.hit.edu.cn

Abstract

We present a bidirectional Example-Based Machine Translation (EBMT) system for Chinese—English. The prerequisite is a bilingual aligned corpus of Chinese—English sentences, and we describe the example extraction efforts purely based on word alignment. The whole system is designed to be language independent and as automatic as possible for construction. We present initial experiments which show that our algorithm can successfully generate better translations for the domain in question than the baseline rule based system.

1. Introduction

The Olympic Games will be held in Beijing, China in 2008. It is clear that a great deal of translation will be required from Chinese to English, and vice versa. This paper describes our bi-directional Example-Based Machine Translation (EBMT) system for Chinese—English, which is purely based on the word alignment information of a given bilingual corpus.

Among the vast issues for a translation system, the followings are emphasized in our system design:

(1) Automatic construction: Manual knowledge composition is not desired for system building except for some public existed knowledge basses (e.g. dictionary). The whole process of translation knowledge acquisition should be as automatic as possible.

(2) Sub-sentential translation example focus: Linguistically there are infinite sentences. A translation system is desired to capture the translation correspondences under the sentence level, hoping to recombine them for proper translations.

(3) Linguistic light approach: Current deep linguistic analysis tools (like parser et al) are not reliable enough. So, if necessary, we would just choose shallow linguistic analyser which causes somewhat less information loss.

(4) Adaptability: Since Olympics demands multi languages besides English and Chinese, the method is kept language independent as possible so that the system success (if it did!) could be readily extended to the translations between Chinese and other languages. For the same reason, domain specific advantages are not exploited in the the system.

2. Auto Word Alignment Based EBMT

To meet the concerns mentioned above, we propose an EBMT method based solely on the word alignment information of the Chinese English parallel corpus. Basically speaking, the whole process of system construction can be fully automatic as long as a translation dictionary and a word aligned bilingual corpus is provided. Figure 1 describes the architecture of the whole system.

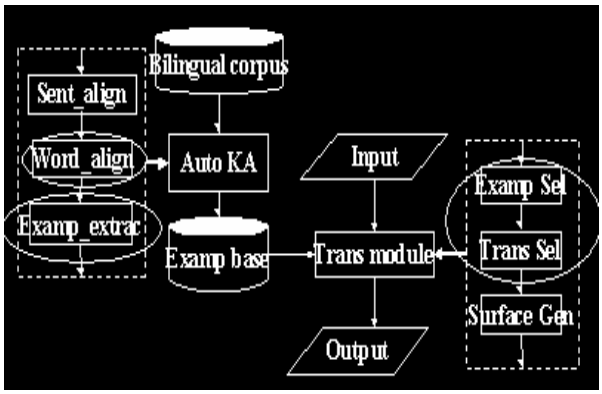

Figure 1: AWA based EBMT architecture.

Roughly speaking, the whole system can be understood as tow parts: the training process (left half of Fig.1) and the translation process (right half).

The training process automatically extracts the translation example base from a word aligned Chinese English bilingual corpus. The key component is the word alignment based example extraction algorithm.

The translation process will pre-process the input sentence (tokenization, numerical processing, Chinese word segmentation et al), feed it into the translation module and display the output. The translation engine will search the best translation via example selection, translation disambiguation and surface generation.

The following of this section will introduce the main parts of the system in detail.

* Supported by the High Technology Research and Development Program of China (2002AA117010-09) and National Natural Science Foundation of China (60375019)

2.1. Word Alignment Based Example Extraction

As mentioned above, current deep linguistic analysis tools are not reliable enough. So we just employ the word correspondence, which is indispensable and somewhat reliable, to formalize the example extraction heuristic.

As shown in figure 2, there are 3 kinds of translation examples recognized by our system: atomic example, extended parallel example and locked example (see figure 2).

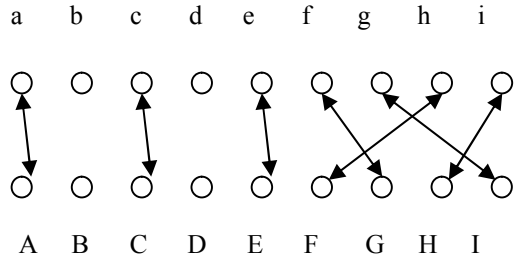

Figure 1: Word link based example extraction.

- Atomic example: just the word correspondences, i.e. (a-A), (c-C), (e-E), (f-G), (g-I), (h-F), (i- H)
- Extended parallel example: combinations of parallel atomic examples with the preceding or following words which is not aligned, i.e.（ab-AB) (bc-BC) (bcd-BCD) (cd-CD) (de-DE). In this case, atomic examples cannot be crossed by another link.
- Locked example: the minimal crossed word links like (fghi-FGHI). Or rather the translation examples in a sentence other than atomic and extended parallel ones.

Such heuristic is purely based on the position pattern of bilingual word correspondence and, consequently, language independent. Although it seems that all words in an aligned sentence pair are utilized, there are fair chances that the extended examples are just noises and locked example is a whole sentence.

2.2. Finding Right Examples for Translation

After getting the translation examples, the EBMT approach for translating a sentence includes 1) find the proper examples, and 2) translation disambiguation.

As for the step 1, our system adopts the dynamic programming to find all possible sequences of translation example combination for the source sentences. This is just like the process of Chinese word segmentation: with translation examples as the Chinese words and the source sentence to be segmented by the eamples.

Suppose one of such sequence S contains l segments:

$$S = [s_{k_0} ..s_{k_1}]^1 [s_{k_1+1} ...s_{k_2}]^2 ...[s_{k_{i-1}+1} ...s_{k_i}]^l \quad (1)$$

where $\{s_i\}$ is the translation example; and a segment is defined as concatenated translations examples from same sentence.

In order to evaluate the properness of a segment, following issues are considered in the system:

- Segment length: Bigger context provides more fixed meaning, thus longer segment is preferred.
- Translation example length: Similarly a segment is desired to be made of longer translation examples.
- Word links: More word alignments, better translation quality the segment has.
- Frequency: a segment is more safe to use if it appears much often in the corpus.

So far the evaluation function for the segment i is designed as:

$$\begin{aligned} &\delta([s_{k_{i-1}+1} ...s_{k_i}]^i) = \\ &(Length([s_{k_{i-1}+1} ...s_{k_i}]^i))^4 \\ &\times An * (1 - \frac{k_i - k_{i-1} + 1}{Length([s_{k_{i-1}+1} ...s_{k_i}]^i)}) \\ &\times \log(\sqrt{Fre([s_{k_{i-1}+1} ...s_{k_i}]^i)} + 1) \end{aligned} \quad (2)$$

where A_n is number of aligned words; *Length(*)* is the segment length; Fre(*) stands for the frequency.

The best segment sequence is just the one with the highest sum of δ scores.

2.3. Translation Disambiguation

As for Step 2, it is necessary if a translation example have multi translations. [Zhanyi, 2002] designed the following formula to search for the best translation sequence among the candidates:

$$T = \arg\max_{T'} P(T' \mid S) * P(An \mid m, l) \quad (3)$$

where,

$P(T' \mid S)$ is a word translation probability model to guarante the reliability of the translation;

$P(An \mid m, l)$ is designed to keep the noise translation away by punishing less word links, in which An is the number of word alignments, m and l is the length of the example and translation respectively; $P(An \mid m, l)$ can be directedly calculated by maximum likelyhood estimation after translation example extraction.

It should be noted here that, to simplify the calculation, techniques like language model and reordering is not considered in current method.

3. Experiments and Performance

To implement the system, we use the algorithm of [Tiejun et al, 2001] to deal with Chinese word segmentation problem. Also a Chinese-English machine translation dictionary with 88,378 entries is used.

The word alignment is processed by a tool described in [Yajuan et al, 2001]. In brief, the method takes into account the translation dictionary information, word similarity, and statistical information to estimate the word correspondence. It can produce more than 80% on F-measure for both general and computer domain bilingual corpus of Chinese and English.

The rival system is a rule-based Chinese-English machine translation system, which is developed by our lab in 2000 and further improved in 2003. Currently the system has 4,000 translation rules manually crafted by translators. And it uses the same Chinese word segmentation algorithm and translation dictionary as EBMT.

The experiments are carried on IWSLT evaluation conditions. IWSLT provided a Chinese English bilingual corpus of basic travel domain, with 20,000 sentence pairs for training. The first result is on the development corpus of the same domain, with 506 sentences and 15 translation references. Both EBMT score and the rival ruel-based MT are reoprted. The second result is the final submission score of EBMT for IWSLT. Note that the rival system is not adapted to the provided corpus. In contrast, the EBMT system is trained on the provided corpus.

Table 1 and Table 2 list the results of 2 tests.

Table 1: IWSLT Development Corpus Score.

		BLEU-4	NIST-5
EBMT	Supplied -Optimal	0.2082	5.5754
	Supplied - Baseline	0.2052	5.3975
	Un-restricted -Optimal	0.2209	5.5940
	Un-restricted - Baseline	0.2236	5.6220
	RBMT	0.1477	5.1990

Table 2: IWSLT Submission Score of EBMT.

	Supplied		Un-restricted	
	Optimal	Baseline	Optimal	Baseline
BLEU 4	0.2099	0.2113	0.2438	0.2427
NIST5	5.9554	5.927	6.1354	6.0603
GTM	0.6013	0.5988	0.6119	0.6152
WER	0.6169	0.6112	0.5941	0.5906
PER	0.5003	0.4976	0.4872	0.4820

4. Discussions and Further Work

From the experiment, we can see that the proposed word alignment based EBMT approach works pretty well for Chinese-English Machine translation. It consistently out-performs the rival rule based system. The simple word alignment based example extraction heuristic possesses the ability to capture the sub-sentential translation correspondence.

We are encouraged by these results. But it should be noted that current performance of our EBMT system yields to the-state-of-art statistical MT. Many defects are found with the system: lack of English inflection generation, insufficient reordering and arbitrary formula integration. It is rather a naive prototype MT system even in the sense of EBMT!

So far the further development work planned includes:

- Validate example extraction heuristic, analyzing its advantages, disadvantages as well as its adaptability to other language pairs;
- Investigate the function of linguistic knowldeg in translation example extraction, checking to if syntax knowledge would help;
- Carry out further experiments on translation module, examining the examples selection and translation disambiguation model;
- Improve English inflection processing by integrating language model or other techniques.

Finally, we are fully aware that this EBMT system could be seed into other MT to form a hybrid engine. We would like to test it in our existed rule based MT and the statistical MT coming into being.

5. References

[1] Liu Zhanyi. 2002. Research and Realization of the Example Based Machine Translation. MSc. Thesis, Harbin Institute of Technology, Harbin, China (in Chinese).

[2] ZhaoTiejun, Lü Yajuan, Yang Muyun and Yu Hao. 2001. Increasing Accuracy of Chinese Segmentation with Strategy of Multi-step Processing. *Journal of Chinese Information Processing* (Chinese Version). (1):13-18

[3] Lü Yajuan, Zhao Tiejun, Li Sheng and Yang Muyun. 2001. Word alignment of Bilingual Corpus Base on Statistic and Lexicon knowledge. In *Proceedings of National Joint Symposium on Computational Lingistics*, Taiyuan, China, pp.108—115.