Make Word Sense Disambiguation in EBMT Practical

Feiliang REN Tianshun YAO

School of Information Science & Engineering, Northeastern University, Shenyang 110004, China
renfeiliang@ise.neu.edu.cn

Abstract. In an EBMT system, we will meet the word sense disambiguation problem. The disambiguation methods used at present can't be used easily in EBMT. We propose a new method for word sense disambiguation in EBMT: it is based on a language mode of the target language. Its main idea is that a proper word sense can make the whole sentence fluent. We use a language model to measure this fluency, and use dynamic programming method to compute the proper words sense sequence in EBMT. It has the virtues of easily being used, and doesn't need extra lingual knowledge, besides, the general performance of it can be improved by using more target language resource to train. And experiment shows our method works well.

Keywords: EBMT, word sense disambiguation, language model, N-gram, dynamic programming

1. Introduction

Example based machine translation (EBMT[1]) is a method of translation by the principle of analogy. When given an input sentence, the EBMT system first retrieval the bilingual aligned corpora and select one or some example sentences as translation templates whose source sentence parts are similar to the given input sentences. And system then modifies the target parts of these translation templates to get finial translation result. These modification operations can be classified as: delete, replace, and insert.

There are many methods for retrieving some translation templates whose source parts are similar to the input sentences. When we have finished this step and begin to the second step: modify the translation templates, the ambiguity problem comes. For example, for a Chinese-Japanese EBMT system, the input Chinese sentence is $S = c_1c_2c_3c_4c_5...c_m$ and what's we retrieved translation template is: $c_1c_2c_4c_5...c_n \leftrightarrow j_1j_2j_3j_4...j_l$, we can see there is a Chinese word c_3 doesn't exist in the source part of the translation template, to get the right translation of the input sentence, we must insert the sense of c_3 in a proper place of the target part of the translation template (the proper insert place can be gotten by comparing two string). If from a bilingual dictionary we know, there are k different senses for the Chinese word c_3, which sense should we choose? The same situation will occur also in replacement operation. This is the problem of word sense disambiguation EBMT.

Traditional word sense disambiguation methods don't work here. They will neither meet a knowledge bottleneck [3-5], nor lower performance [2]. And it can't satisfy the requirement in an EBMT system.

Here we propose a new word sense disambiguation method, which is based on the N-gram language model. We think a proper word sense must be the sense that makes the whole sentence looks frequent, and we use N-gram language model[6,7] to measure this fluency. That is to say in our disambiguation method, we select the word sense that make the whole sentence fluent most.

Our paper is organized as following: in section 2, we give the detail description of our disambiguation method, and its computation algorithm; in section 3, are our experiments; and at last, in section 4, we drew our conclusions.

2. Our Disambiguation Method

Suppose we need to insert ***m*** words' sense in the target part of a translation template, and every word has ***n*** different senses. Our aim is to select a proper sense sequence to be inserted that can make the target part of the translation template fluent most. Let's suppose j_i is a word in the word sequence of the target part of a translation template. P_i is the place where a new word should be inserted. s_{ij} is the j-th sense of the i-th word to be inserted. And the disambiguation in EBMT can be shown in figure 1.Its task is to find a sense path that makes the whole sentence fluent most.

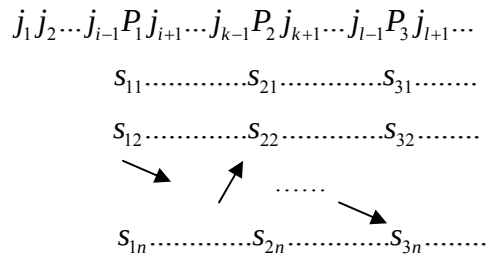

Fig. 1. Word Sense Disambiguation in EBMT

Our disambiguation method can be denoted as the following formula:

$$BestSequence(s_{ij}) = \arg\max P(S) \quad i = 1 \to m,\ j \in [1, n] \tag{1}$$

Where $P(S)$ is the probability of sentence, and we use N-gram language model to compute it. Suppose we use trigram language model, then $P(S)$ can be computed as following:

$$P(S) = P(j_1)P(j_2 \mid j_1)P(j_3 \mid j_1 j_2) \cdots P(j_n \mid j_{n-2} j_{n-1}) \tag{2}$$

To make the time cost lower, we use dynamic programming method.

Let $\delta_t(i)$ denotes the max probability for the current sentence when take the i-th sense at place t, we can write $\delta_t(i)$ as follow:

$$\delta_t(i) = \max P(j_1, \cdots, j_{t-1}, j_t = s_{ti} \mid S) \quad (3)$$

Let $\Delta_t(i)$ denotes the word sense we choose at place t-1 is i. That is to say we use the variable $\Delta_t(i)$ to write down the path of the word sense selection sequence. Our disambiguation computation algorithm is in figure 2.

Initialize: $\delta_1(i) = P(s_{1i} \mid j_{1-1} j_{1-2})$

$\Delta_1(i) = 0$ (j_{1-1} and j_{1-2} are the two previous words of the word s_{1i})

Recursion: $\delta_t(i) = \max \delta_{t-1}(i) P(s_{tj} \mid j_{t-1} j_{t-2})$ $1 \le j, i \le n$

$\Delta_{t-1}(j) = j$

Calculation: $P(S) = \arg\max \delta_m(i)$

Trace back: $\bar{X}_T = \arg\max[\delta_T(i)]$

$\bar{X}_t = \Delta_{t+1}(\bar{X}_{t+1})$ ($\bar{X}_t$ denotes the inserted word sense at place t.)

Fig. 2. Our computation algorithm for disambiguation

Our disambiguation algorithm looks like the Viterbi algorithm, and its time complicity is $O(n^2)$.

3. Experiments

We collected 1,400,000 words' Japanese monolingual corpora from the Internet for the training of language model. In table 1 we give some basic information of the corpora. We use trigram language model for disambiguation. And we select 2,000 Chinese sentences as input for translation. We use the correct ratio to evaluate the performance of our disambiguation method. Denotes N as the word number that should be inserted when translating. Denotes N_c as the word number that has been correctly selected for sense. We denote the correct ration for disambiguation as following formula:

$$CorrectRatio = (N_c / N) \times 100\% \quad (4)$$

Our experiments results are in table 2. From the experiments results, it seems our disambiguation doesn't show much power, but you know, we only use about 37,000 Japanese sentences to training the

language model, and if we increase the sentences of training corpora, the performance will increase accordingly. In our experiments, most of the errors for the disambiguation task are because of the data sparsely. We can improve the performance by increasing the scale of the training corpora.

Table 1. Basic information of teh corpora used for training language model

Sentence number	Average words per sentence	Average insert and replace words
36962	39.7	1.7

Table 2. Experiments result for our disambiguation method

Input sentence number	Average insert words per sentence	Correct ratio
2,000	2.3	77.2%

4. Conclusions

In this paper, we propose a new disambiguation method for EBMT based on N-gram language model. Its main idea is to try to search a word sense sequence that can make the whole sentence fluent most. It doesn't need large number of disambiguated training data, and is easily used. And experiments show our disambiguation method works well in a Chinese-to-Japanese EBMT system.

References

[1]Harold Somers.2001. Review Article: Example-based Machine Translation. Machine Translation 14, pp.113-157

[2]Christopher D.Manning, Hinrich Schutze. 2005. Foundations of Statistical Natural Language Processing[M]. pp.143-163

[3]Gale, William A, Kenneth W.Church, David Yarowsky, 1992b. A method for disambiguating word senses in a large corpus[J]. Computers and Humanities 26: 415-439

[4]Brown, Peter F, Stephen A, et al. 1991b. Word sense disambiguation using statistical methods. ACL 29, pp.264-270

[5]Black, Ezra. 1988. An experiment in computational discrimination of English word sense. IBM Journal of Research and Development 32:185-194

[6]Eiji Aramaki, Sadao Kurohashi, Hideki Kashioka, Hideki Tanaka:Probabilistic Model for Example-based Machine Translation, MT Summit X, pp.219-226, 2005.

[7]Michael Carl and Andy Way (editors), Recent Advances in Example-Based Machine Translation, Kluwer Academic Publishers, 2003.