Similarity Comparison between Chinese Sentences

Lina Zhou[12]   James Liu[2]

[1]Institute of Computational Linguistics, Peking University
Beijing, China , 100871
cslzhou@comp.polyu.edu.hk

[2]Department of Computing, Hong Kong Polytechnic University
Kowloon, Hong Kong
csnkliu@comp.polyu.edu.hk

Abstract

Identifying and extracting similar sentences from the example base is an essential procedure in machine-aided human translation (MAHT) and example-based machine translation (EBMT) system. A method for measuring the similarity between a pair of Chinese sentences has been proposed in this paper. Obviating from the common thesaurus-based strategy, a new principle based on word grammatical features is presented thereafter. Moreover, a dynamic mechanism is built into the method to increase the robustness and flexibility of the matching algorithm. From observations on the initial results, we' ve found that the expected most similar sentence in the example base for an input is listed among the first four candidate sentences in most cases, which is very helpful for both MAHT and EBMT.

1. Introduction

It is regarded as an essential process to measure the similarity between an input sentence and the stored examples or translation candidates in machine-aided human translation system and example-based machine translation system.

There has been no accurate definition for similarity comparison available in the field of machine translation, though, it is clear that similarity comparison is a cloning process, which measures the matching scores between two objects in terms of certain similarity metric. As far as sentence pairs are concerned, the purpose of similarity comparison is to identify and extract sentences from the stored base, which are similar to the input sentence. The criteria for comparison can be based on attributes of the word, phrase structure or sentence. The most popular strategy is based on the thesaurus relation. (Furuse, 1992; Nirenburg, 1993; Sato, 1992; Cranias et. al., 1994; Maruyama, 1992; Zhang et. al. 1995). Either the performance of such method is not satisfactory or it relies much on pre-processing efforts to obtain acceptable results. As a result of trade-off among the factors influencing the accuracy and efficiency of the matching algorithm, a word grammatical attribute oriented approach is proposed for comparing Chinese sentences, which takes the following significance:

- It conforms with the most frequently used syntax-based translation techniques. The basis upon which similarity metric is built can be directly applied in the transfer stage.
- It explores the prospect of word feature oriented approach and the possibility of improving comparison results by elaborating the grammatical features of words.

In this paper, the knowledge base for the similarity metric will be presented in the next section. The new metric for measuring similarity will be described in Section 3 and followed with an algorithm in Section 4. Finally, some experimental samples are given in Section 5.

2. Grammatical Knowledge-base

The knowledge base was originated from the *Electronic Dictionary of Grammatical Information for Contemporary Chinese* (Yu et. al., 1996). Since the dictionary was designed for general-purpose applications, elaborately defined features have to be filtered or selected to facilitate the identification of the right translation candidates when applied to sentence comparison. Following the above principles, eight sub-dictionaries were employed, i.e. noun, verb, adjective, adverb, pronoun, classifier, preposition and time dictionaries, etc. The specific features helpful for sentence comparison were selected in every dictionary.

3. Similarity Metric

Based on the features defined for each category, the similarity metric between a pair of Chinese sentences (A,B) was defined as:

Assume that A = $a_1a_2......a_n$, B = $b_1b_2......b_m$, $a_i(b_j), 0<i<n+1, (0<j<m+1)$ is the *ith* (*jth*) word in sentence A (B). F is the whole feature set of a certain word category, E a subset of F, and |E| stands for the number of features in E. *fea$_k$(a), sub_pos(a) and pos(a)* represent the *kth* feature, sub-category and part-of-speech of word *a* respectively. *Ss(A,B)* represents the similarity metric between A and B, while *$S_w(a_i,b_j)$* the similarity score between a_i and b_j. $a_1^i(b_1^j)$ represents the string from a_1 (b_1) to a_i(b_j). L(A,B) is the normalizer for the sum of the similarity score.

$$Ss(A,B)=\frac{Ss(a_1^n,b_1^m)}{L(A,B)} \tag{1}$$

$$Ss(a_1^i,b_1^j)=\begin{cases}0,\ if\ i<1\cup j<1\\ Ss(a_1^{i-1},b_1^{j-1})+Sw(a_i,b_j),\ if\ i>1\cap j>1\cap Sw(a_i,b_j)>0\\ Ss(a_1^i,b_1^{j-k}),\ else\ if\ j>k>1\cap Sw(a_i,b_{j-k})>0\\ Ss(a_1^{i-1},b_1^j),\ otherwise\end{cases} \tag{2}$$

$$Sw(a_i,b_j)=\begin{cases}0 & if\ pos(a_i)\neq pos(b_j)\\ 0.25 & else\ if\ sub\_pos(a_i)\neq sub\_pos(b_j)\\ 0.5 & else\ if\ \bigcup_{\substack{k\in E\\ E\subset F\\ |E|\leq 0.5*|F|}} fea_k(a_i)=fea_k(b_j)\\ 0.8 & else\ if\ \bigcup_{\substack{k\in E\\ E\subset F\\ 0.5*|F|<|E|<|F|}} fea_k(a_i)=fea_k(b_j)\\ 1 & else\end{cases} \tag{3}$$

In contrast with the common static definition for *L(A,B)*, a new and dynamic formula is given thereafter:

$$L(A,B)=N(n,m)+F(n,m)/3 \tag{4}$$

where *N(n,m)* is the number of comparison times; while *F(n,m)* is the number of words failed to get matched. It could be in some cases that the algorithm doesn' t provide the optimum matching for an input sentence. However, the adoption of a dynamic mechanism does ensure better efficiency for matching, and *F(n,m)* is introduced as a penalty factor to improve the results.

4 Algorithm

For an input sentence A= $a_1a_2......a_{n+1}$ [1]and a stored sentence B=$b_1b_2......b_{m+1}$,

[1] Here an extra word is appended to indicate the end of the sentence.

```
1. Initialize $i = 1, j = 1; t = 0, f = 0;$
2. While $a_i \neq a_{n+1}$
       if $b_j \neq b_{m+1}$
           if $S_w(a_i, b_j) < 0.25$
               $j_0 = j$
           else
               $i = i + 1$
               $Sum = Sum + S_w(a_i, b_j)$
           endif
           $j = j + 1$
           $t = t + 1$
       else
           $j = j_0$
           $f = f + 1$
           $i = i + 1$
       endif
4. $S_s(A,B) = \frac{Sum}{t + f/3}$
```

5. Experimental Samples

A parallel bilingual corpus with about 3,000 Chinese and English sentence pairs has been utilized and pre-processed (Zhou and Liu, 1997). Both sides have been annotated with part-of-speech.

The testing results are classified into five categories: complete match, word replacement, word insertion and deletion, phrase replacement and modification, and composition, etc. Several samples are provided for explanation:

(1) Word Replacement

In: 東京是世界上　人口最多的城市[2]·

Re: 上海/n 是/v 世界/n 上/f 最/d 大/a 的/u 城市/n 之一/m ·/w (Shanghai is among the largest cities in the world.) <0.85>

[2] The focus part in each category is underlined. The similarity score of the result (Re) for the input (In) is put in the brackets.

(2) Phrase Replacement and Modification

In: 我有足夠的錢買 書・

Re: 我/r 有/v 足夠/v 的/u 錢/n 買/v 這/r 兩/m 本/q 書/n ・/w (I have enough money to buy these two books.) <0.89>

5. Conclusion and Future Directions

A new algorithm for measuring the similarity between a pair of Chinese sentences has been proposed in this paper. It emphasizes the grammatical features of Chinese words supported by the comprehensive electronic dictionaries. In addition, a dynamic mechanism and penalty score are built in to increase the robustness and flexibility of the algorithm. From observation on the matching results, we feel that most of the selected sentences are much related with the input on the syntactic and even semantic level. The expected most similar sentence from the example base is listed among the first four candidate sentence s in most cases, which is considered to be very informative and helpful for both MAHT and EBMT. In view of the basic ideas introduced, the approach is easy to be tested on other languages.

References

Furuse, O. and H. Iida, "An Example-based Method for Transfer-driven MT", in TMI' 92, , 1992, pp. 139-148.

Nirenburg, S. "Two Approaches to Matching in EBMT", in TMI'93 , 1993, pp. 47-57.

Sato, S. "CTM: An Example-based Translation Aid System", in COLING'92 , 1992, pp. 1259-1263.

Cranias, L. et. al., "A Matching Technique in Example-based Machine Translation", in COLING' 94, 1994, pp.100-104.

Maruyama, H. "Tree Cover Search Algorithm for EB Translation", in TMI'92. 1992, pp. 173-184.

YU, S.W., Y.F. Zhu, H. Wang and Y.Y. Zhang. "Specification for Grammatical Information Dictionary of Contemporary Chinese", in Journal of Chinese Information Processing, Vol. 10,No.2, 1996, pp.1-15.

Zhou, L.N. and J. Liu, "Extracting More Word Translation Pairs from Small-sized Bilingual Parallel Corpus: integrating rule and statistics-based method", in ICCPOL'97. 1997, pp.250-255.

Zhang, M., S. Li, T.J. Zhao and M. Zhou. "An Algorithm for Calculating the Similarity between Chinese Sentences and its Application", in the Third National Joint Conference on Computational Linguistics, Tsinghua University Press, 1995, pp. 152-158.