University of Macau

Abstract

CROSS SENTENCE ALIGNMENT BASED ON SINGULAR VALUE DECOMPOSITION

by Anna Ho

Thesis Supervisor: Prof. Li Yi Ping and Dr. Wong Fai
Master of Science in Software Engineering

Many researchers have already contributed a lot in the sentence alignment with parallel corpora. Most of the researches focus on linguistic structurally similar bilingual sentence alignment or English-Chinese sentence alignment methods. Due to the lack of research in Portuguese-Chinese sentence alignment methods, and limited resources provided in parallel translations, we aimed at investigating a sentence alignment framework which accepts both comparable and parallel Portuguese-Chinese corpus. In our case, comparable corpus includes magazines, newspaper, articles, reports, etc. Since aligning linguistic structurally dissimilar comparable corpus is not a trivial task, in this thesis, we provide appropriate solutions to the problems.

In order to solve the problem of generating the basic relationship score between each possible bilingual sentence pair, we make use of the Portuguese-Chinese bilingual dictionary to provide a set of mathematical models which based on the lexicon information provided by the sentence itself. However, sub-phrase problem may occur when a shorter Portuguese sentence is compared with a longer Chinese sentence. To overcome this problem, a penalty mechanism is employed and a relationship coefficient is defined for penalizing the relationship score according to the remaining Chinese characters which cannot be matched.

Since our framework accepts comparable corpus, we have to face the cross sentence alignment problem caused by its nature, which means that the first sentence in one

language may not correspond to the first sentence in the target language. In order to solve the problem, this thesis proposed stop words filtering and the application of Singular Value Decomposition (SVD) technique with dimensionality reduction to the alignment framework. Useless and noisy information which disturb the correct choice of alignment are eliminated. Moreover, since hidden information has been revealed, the true relationship among sentence pairs can be distinct more clearly. Therefore, the cross sentence alignment problem can be solved more fluently when compared with the framework which without the application of SVD.

In order to fully utilize the advantage of SVD, it is very important to provide the least but most adequate statistical relationship information among sentence pairs to the SVD alignment framework. In the next step, we extent the framework to deal with the weaknesses in calculating the statistical relationship score, such as the absence of preserving sequential order of Chinese characters and excluding special alignment clues. Bilingual documents segmentation and N-gram based lexical matching model are introduced to solve the sequential order problem. Moreover, to balance the employment of different alignment clues from different types of bilingual documents, scoring feature algorithm is introduced which includes the following four features: lexical matching, length ratio probability, punctuation matching and digit group matching.

We applied the cosine similarity measure to the analyzed results to retrieve the aligned sentence pairs. In this research, the performance of the proposed algorithm achieves over 80% accuracy in parallel and comparable corpus which shows the effectiveness of our research.