University of Macau

Abstract

Unsupervised Word Sense Disambiguation using non-aligned bilingual corpus in application to Portuguese-Chinese Machine Translation

by Francisco de Oliveira

Thesis Supervisors: Prof. Li Yi Ping and Dr. Wong Fai

Master of Science in Software Engineering

Many words in natural language are known to be highly ambiguous. For example, the noun "Português" (Portuguese) can either be in the sense of a language or a human with a Portuguese nationality.

The objective of Word Sense Disambiguation (WSD) is to identify the correct meaning or sense of a word in a given sentence.

In order to obtain good results in the disambiguation process, usually it requires a set of resources, including a large corpus, a lexicon, an annotated Treebank, and sense inventories. However, not all the languages have all the mentioned resources in a digitized format, especially for the languages that we are in concern, Portuguese and Chinese. Moreover, data sparseness is another common problem in Statistical Corpus based WSD approaches.

To overcome the mentioned problems, this thesis presents an Unsupervised and Non-Aligned Corpus (UNAC) based framework for WSD that relies on an unsupervised learning and a set of bilingual resources, including a non-aligned corpus, a bilingual lexicon, and a sense inventory. The proposed framework first identifies words related to each of the ambiguous words based on their surrounding words, relative distance, and mutual information. The disambiguation algorithm is based on a mathematical model that identifies the most suitable sense of an ambiguous word in

terms of the related words. Moreover, the use of bilingual examples and Singular Value Decomposition techniques are applied to overcome the data sparseness problem. All the senses learned are converted into a set of rules and stored in the database.

This framework is then applied in a Portuguese-Chinese MT System, which is based on the formalism of Constraint-based Synchronous Grammar. In this research, it also applies shallow parsing techniques in the identification of Noun Phrases to further enhance the efficiency and the translation quality of the system.

Empirical results show that the UNAC based framework for WSD is superior to the baseline approach and there is an improvement in the translation accuracy after applying the framework in the Portuguese-Chinese MT System.