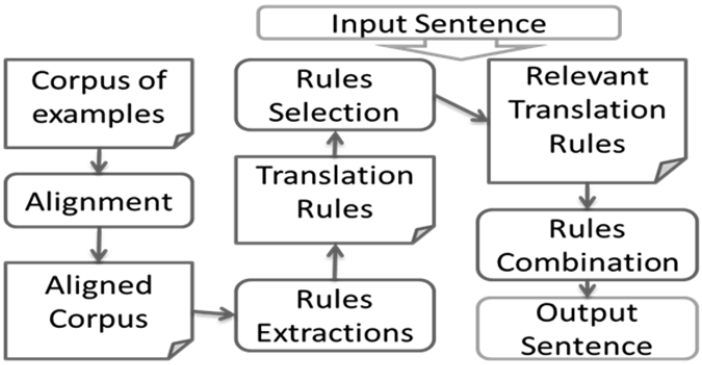
Fabien Cromieres（黒橋教授）

「Using Scalable Run-Time Methods and Syntactic Structure in Corpus-Based Machine Translation」

（スケーラブルな実行時手法と構文木に基づくコーパスベース機械翻訳）

平成 23 年 3 月 23 日授与



With the development of Internet and Globalization, Machine Translation of natural languages has become a more strategic topic than ever. In the past decade, increase in processing power and available resources have led many researchers to shift focus from heavily linguistic and expert knowledge-based systems to corpus-based approaches. The principle of the corpus-based approach is to use existing corpora of translations to automatically learn how to create new translations. This approach started to develop in the 1980s, and it actually originated in Kyoto University with Professor Makoto Nagao, who proposed some of the initial ideas as early as 1984. The general framework for Corpus-Based machine translation is as follows. First, a large corpus of translation examples (consisting of sentences in source and target languages) is collected. These examples are then aligned: links are established between subparts of the source and target side that are translations of each others. The aligned corpus can then be used to extract translation rules that will be used for translating an input sentence. In this general framework, many differences exist between systems, depending on which type of translation rules are used and which amount of syntactic information is used.

This thesis explores several issues concerning use of syntactic structure in large-scale Corpus-Based Machine Translation. In particular, we emphasize the need to move towards methods that are scalable (applicable to systems using millions of examples) and if possible relying on run-time computations (avoiding the need to pre-compute and store too many informations). We also emphasize the need to take into account the syntactic structures of the examples and their differences across language.

More specifically, the contributions of this theses cover several aspects of corpus-based machine translation. As for the alignment aspect, we propose a new framework for alignment making use of graphical models. This framework allows specifying new alignment algorithms in a flexible way. We also propose an approach to alignment where the alignment can be done on a per-sentence basis, allowing to do the alignment step at run-time, on an on-demand basis. As for the rules extraction and selection aspect, we develop a method for efficiently retrieving translation examples in the corpus that are pertinent to the translation of a query sentence. This method is based on retrieving examples whose part of the syntactic structure match that of the sentence. Several challenges, such as keeping the processing time and memory used manageable, require the use of innovative algorithms, such as an adaptation of the concept of Suffix Arrays to trees. Finally, motivated by the facts that some of the simpler type of translations rules cannot capture some translation phenomenon happening during translation, we also propose some new type of rules that can handle more complex transformation phenomenon while still being time-efficient to use.