Chenhui Chu （黒橋教授）
「Integrated Parallel Data Extraction from Comparable Corpora for Statistical Machine Translation」
（統計的機械翻訳におけるコンパラブルコーパスからの対訳データの統合的抽出）
平成 27 年 3 月 23 日授与

Machine translation (MT), as a high level application of natural language processing (NLP), is a powerful tool to improve the efficiency and reduce the cost of translation. Over the last decade or two, statistical machine translation (SMT) has been the main approach in both the research community and the commercial sector. In SMT, translation knowledge is automatically acquired from parallel corpora (sentence-aligned bilingual texts), making the rapid development of MT systems for different language pairs and domains possible once parallel corpora are available. Because of the high dependence on parallel corpora, the quality and quantity of parallel corpora are crucial for SMT. However, except for a few language pairs and some specialized domains, high quality parallel corpora of sufficient size remain a scarce resource. This scarceness of parallel corpora has become the main bottleneck for SMT.

Comparable corpora are a set of monolingual corpora that describe roughly the same topic in different languages, but are not exact translation equivalents of each other. Exploiting comparable corpora for SMT is the key to addressing the scarceness of parallel corpora. The reason for this is that comparable corpora are far more available than parallel corpora, and there is a large amount of parallel data contained in the comparable texts. The main focus of this thesis is extracting the parallel data from comparable corpora to improve SMT. Comparable corpora potentially contain three types of parallel data: bilingual lexicons, parallel sentences and parallel fragments. In this thesis, we propose novel approaches to extract these three types of parallel data from comparable corpora in an integrated framework.

The overview of our framework is presented in Figure 1. As initial, we have comparable corpora and a small seed parallel corpus. We first generate a bilingual dictionary from the seed parallel corpus. As the coverage of this dictionary is low, we further extract bilingual lexicons from comparable corpora ((1) in Figure 1) and combine them with the generated dictionary. Using the combined dictionary, we can apply cross-lingual information retrieval to generate parallel sentence candidates from comparable corpora. Next, we apply parallel sentence extraction that can classify the parallel sentence candidates into parallel and comparable sentences ((2) in Figure 1). Finally, we apply parallel fragment extraction to extract parallel fragments from the comparable sentences ((3) in Figure 1). The combined dictionary can be used for both parallel sentence and fragment extraction. Moreover, the extracted parallel sentences can be used to support parallel fragment extraction. The extracted parallel sentences and fragments are used as training data for SMT. Also, they can be appended to the seed parallel corpus for bootstrapping. Experiments verify the effectiveness of the proposed approaches for the scarceness of parallel corpora that SMT suffers.

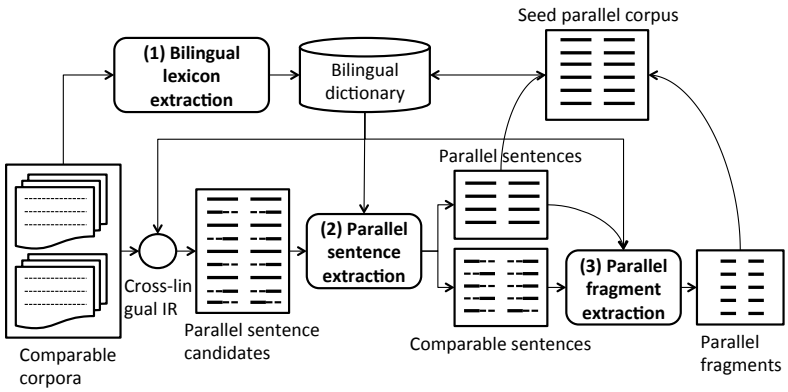

Figure 1: Integrated parallel data extraction framework