# 論文の内容の要旨

論文題目　Statistical Machine Translation Using Large-Scale Lexicon and Deep Syntactic Structures

（大規模辞書及び深い文法構造を用いた統計的機械翻訳）

氏名　呉　　先超（Xianchao WU）

（本文）The goal of this thesis is to establish a Statistical Machine Translation (SMT) system by making use of a large-scale lexicon and well-grained translation rules. A semi-supervised approach is proposed for mining a million word level bilingual lexicon from the Web. A linear-time algorithm is proposed for extracting well-grained translation rules from deep syntactic structures including predicate argument structures, which are generated by a Head-driven Phrase Structure Grammar (HPSG) parser. Extensive experiments on English-to-Japanese and Chinese-to-English translation testified that the proposals improved state-of-the-art.

These proposals are motivated by tackling two primary challenges in building a practical SMT system. One is the difficulty of keeping up with neologism in parallel corpora used for training. For example, new technical terms are frequently emerging in newly published technical papers. Another challenge is the requirement of integrating rich syntactic information (e.g., tense, voice, etc.) and semantic dependencies into a SMT system.

This thesis first proposes *a semi-supervised approach for mining a Chinese-English lexicon* from parenthetical expressions in Chinese corpora. The approach is based on an observation that many Chinese neologisms (such as technical terms, movie names, personal names, etc.) are accompanied by their English translations in parentheses for explaining them. The mining task is to pick a Chinese phrase from the text before the parenthesis such that it exactly corresponds to the English text in the parenthesis. A novel idea in our proposal is to obtain new entries by using a self-trained transliteration/translation model as a classifier to distinguish a good Chinese phrase from bad ones. The experiments testified that our approach achieved higher precision and recall than the previous supervised approach. On both the personal name transliteration task and the technical term translation task, our new mined lexicons do help to significantly improve the translation accuracies. As a result, 1.1 million English-Chinese distinct entries were mined with the accuracy of 77.7% from 252G Chinese Web pages and 55G technical papers.

Then, this thesis uses *several algorithms for utilizing deep syntactic analysis in syntax-based SMT*

*models*. In order to utilize the tree structure of HPSG, the linear-time GHKM algorithm for extracting CFG translation rules is adopted to include typed feature structures. Furthermore, to utilize the semantic representation given by the HPSG parser, we identify sub-trees in a parse tree that correspond to basic units of the semantics, namely a sub-tree covering a predicate and its arguments. We propose a linear-time algorithm for identifying the sub-trees to be used in tree-to-string translation rules.

These tree-to-string translation rules are employed in two SMT models. First, *a tree-to-string translation model* is built for English-to-foreign language translation. The decoding algorithm is based on a bottom-up traversal of the source tree, with translation candidates attached to each tree node. Second, *a string-to-tree translation model* is constructed for foreign language-to-English translation. In this model, the tree-to-string translation rules are used in an inverse direction. The decoding procedure is to construct a target language parse forest through parsing a source sentence. The problem during decoding is to maintain cubic time of the length of the source sentence even when complex data structures such as typed feature structures and predicate argument structures are included in the translation rules. Dealing with this problem, an existing linear-time algorithm is adopted to inversely binarize the HPSG tree-to-string rules into Chomsky Normal Form to be used with an n-gram language model integrated in CKY decoding.

Finally, this thesis describes extensive experiments on large-scale parallel corpora to evaluate the translation models. Compared with state-of-the-art hierarchical phrase-based and our implementations of CFG-based SMT systems, the models described in this thesis achieved significant improvements on BLEU score for English-to-Japanese and Chinese-to-English translations. Further improvement was achieved when utilizing the large-scale lexicon for Chinese-to-English translation.