UNIVERSITY OF AMSTERDAM

Aligning the Foundations of Hierarchical Statistical Machine Translation

*G.E. Maillette De Buij Wennige*

Gideon Maillette de Buy Wenniger

==========

Abstract

Statistical machine translation (SMT) plays an important role in the automatic translation of the large
and increasing volume of documents that has become globally available.
The results of SMT are often still lacking in various aspects including word order.
This thesis focuses on the improvement of hierarchical SMT, in particular *Hiero*.
*Hiero* rules lack nonterminal labels. This gives them little context and makes
their combination into full translations poorly coordinated, and strongly dependent on the language model.

In this thesis, bilingual labels are added to *Hiero* rules.
These bilingual labels lead to more coherent translations with better word order, as demonstrated by extensive experiments on three language pairs.
The proposed labels require no syntactic information, and use only the information from word alignments.
This distinguishes them from various types of syntactic labels earlier proposed in the literature.
Bilingual labels are based on a newly proposed framework called hierarchical alignment trees (HATs).
HATs are bilingual trees that represent the hierarchical translation equivalence structure induced from word alignments.
HATs maximally decompose word alignments into phrase pairs, and provide an explicit description of
the local reordering taking place within each phrase pair.

The last part of the thesis is concerned with the complexity of empirical translation equivalence.
Given a word alignment and a grammar, it studies the question what it means for the grammar to cover the word alignment.
HATs play a key role in answering this question exactly and efficiently, and
are applied to characterize alignment complexity for various language pairs.