This dissertation deals with the improvement of systems for machine translation (MT) using semantic information. Such information tends to remain constant during translation, while the syntactic structure of sentences often changes, as a result of linguistic necessities or translators' choices. These changes make it difficult to derive syntactic rules automatically when building a statistical MT system (a type of data-driven system) using a substantial amount of sentences and their translation. For instance, the verb in a subordinaute clause must be moved after the direct object when translating from English to Dutch. Another example relates to the verb like: when translating it to bevallen ('please') in Dutch, the direct object becomes the subject. Constructing a syntax-based statistical MT system involves the automated alignment of words, the creation of a phrase table with the translation of words and word groups, and the derivation of translation rules based on syntactic trees produced by a parser.

In this dissertation, we investigate whether a semantic analysis of sentences and their translation facilitates the creation of translation rules and improves the quality of rules. We focus on shallow semantics, in the form of predicates and roles, and experiment with a four-step approach which requires a minimum amount of manual intervention. The first step consists of enriching parse trees with predicate and role labels. As tools which perform such labeling are scarce, we design a method which supports the creation of a new tool on the basis of semantic information in another language. This method makes use of word alignment and creates mappings between syntax and semantics. The second step consists of aligning parse trees via semantic labels. The third step consists of deriving translation rules based on semantic alignment. The final step extends a statistical MT system with semantic translation rules.

We implemented our four-step approach in order to evaluate it. The results indicate that enriching parse trees with semantic predicate and role labels leads to more precise tree alignment results, and that combining a phrase table with semantic translation rules helps in improving translation quality. While we perform tests on the language pair English-to-Dutch, our approach is sufficiently generic for tests on other language pairs and for contexts other than MT. For instance, it can be applied for detecting specific structures in aligned parse trees in the context of translation studies.

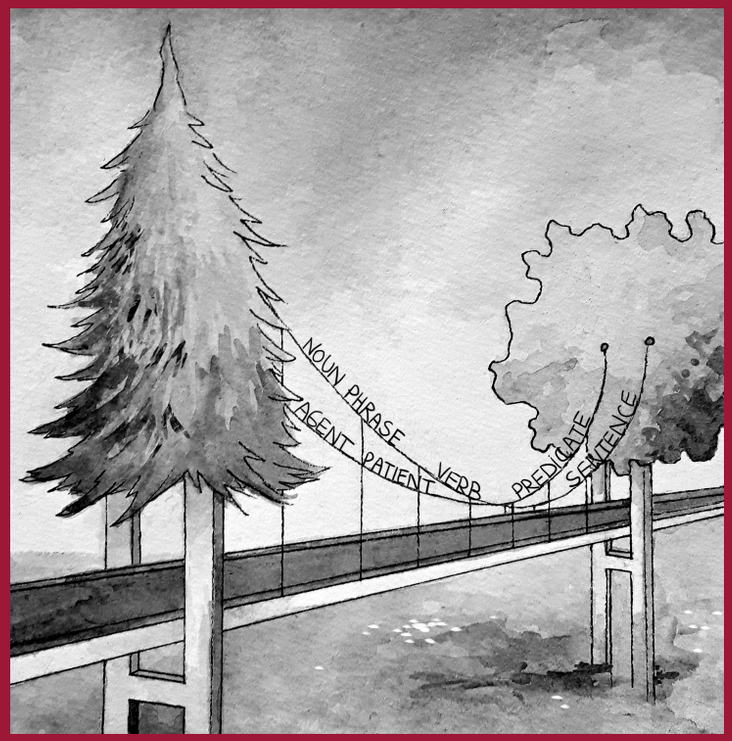
: LOT
Netherlands Graduate School of Linguistics
ISBN 978-94-6093-275-5

: LOT 490

Tom Vanallemeersch
Data-driven Machine Translation using Semantic Tree Alignment

Tom Vanallemeersch

Data-driven Machine Translation using Semantic Tree Alignment



Department of Linguistics

: LOT
Netherlands Graduate School of Linguistics
Landelijke Onderzoekschool Taalwetenschap