Abstract

Bimorphism Machine Translation
Daniel Quernheim
(Dissertation, 2017)

The field of statistical machine translation has made tremendous progress due to the rise of statistical methods, making it possible to obtain a translation system automatically from a bilingual collection of text. Some approaches do not even need any kind of linguistic annotation, and can infer translation rules from raw, unannotated data. However, most state-of-the art systems do linguistic structure little justice, and moreover many approaches that have been put forward use ad-hoc formalisms and algorithms. This inevitably leads to duplication of effort, and a separation between theoretical researchers and practitioners. In order to remedy the lack of motivation and rigor, the contributions of this dissertation are threefold:

1. After laying out the historical background and context, as well as the mathematical and linguistic foundations, a rigorous algebraic model of machine translation is put forward. We use regular tree grammars and bimorphisms as the backbone, introducing a modular architecture that allows different input and output formalisms.

2. The challenges of implementing this bimorphism-based model in a machine translation toolkit are then described, explaining in detail the algorithms used for the core components.

3. Finally, experiments where the toolkit is applied on real-world data and used for diagnostic purposes are described. We discuss how we use exact decoding to reason about search errors and model errors in a popular machine translation toolkit, and we compare output formalisms of different generative capacity.

Keywords

computational linguistics, machine translation, bimorphisms, synchronous grammars, tree automata, formal language theory