Statistical MT Systems Revisited: How much Hybridity do they have?

Hermann Ney
RWTH Aachen University, Aachen and DIGITEO Chair, LIMSI-CNRS, Paris
Lehrstuhl für Informatik 6
RWTH Aachen
Ahornstr. 55
52056 Aachen
ney@informatik.rwth-aachen.de

Abstract

The statistical approach to MT started about twenty-five years ago and has now been widely accepted as an alternative to the classical approach with manually designed rules. Among the attractive properties of the statistical approach is its capability to learn the translation models automatically from a (sufficiently) large amount of source-target sentence pairs. Thus the need for the manual design of suitable rules and for human interaction can be reduced dramatically when developing an MT system for a new application or language pair.

The idea of hybrid MT is to combine the advantages of both the rule-based and statistical approaches. In practice, most statistical MT systems make use of manually designed rules in order to improve the MT accuracy. We revisit the RWTH systems in order to study the effect of typical preprocessing steps based on manually designed rules. The RWTH systems cover various tasks (e.g. news, patents, lectures) and various languages (e.g. Arabic, Chinese, English, Japanese). The preprocessing steps may include a categorization of numbers, date and time expressions, a word decomposition based on morphological analysis and explicit word re-ordering based on a syntactic analysis. In general, the preprocessing steps may depend heavily on the language pair under consideration.

We will also address concepts that aim at a tighter integration of the conventional rule-based and the statistical approaches. We will consider the implications of such a tight integration for the architecture of an MT system.

Proceedings of the Second Workshop on Hybrid Approaches to Translation, page 7,
Sofia, Bulgaria, August 8, 2013. ©2013 Association for Computational Linguistics