

Machine Translation: Green, Yellow, and Red

Aarne Ranta, University of Gothenburg

Abstract

Machine translation (MT) can be divided into quality-oriented and coverage-oriented approaches (also known as *dissemination* and *assimilation*, respectively). The current main stream is coverage-oriented: most people use MT for getting an idea of what some document is about, but don't rely on it when they want to publish their own documents. Coverage-oriented systems must be able to translate everything, whereas quality-oriented systems usually have to sacrifice coverage and specialize on some domain.

Most available coverage-oriented systems are statistical (Google translate, Bing), but there are also rule-based systems available (Systran, Apertium). In MT research, the main line of research is hybrid systems combining statistics with linguistic knowledge. In this talk, we will present a hybrid MT approach based on GF, Grammatical Framework.

Most of the previous work in GF has focused on small, quality-oriented systems working on controlled languages; the main asset has been the scalability to high numbers of parallel languages. But recent developments in GF runtime algorithms and language resources have made it possible to address the coverage-oriented task of "translating everything". This happens of course with some loss of quality, but the great advantage of GF (and some other knowledge-based systems) is that we can make a clear distinction between levels of confidence. We have used this knowledge in translation programs by marking translations as green (reliable), yellow (grammatically correct but unreliable), and red (unreliable but "still better than nothing"). There is also a clear recipe for improving the quality by increasing the size of the "green" area. These are typically inherited from domain-specific controlled languages.

The talk will explain how grammars of different levels are created and combined, how statistics is used in the translation process and for bootstrapping grammars, and how the resulting system performs in comparative evaluation.

The current system is available in eleven languages, both as a web service and as a mobile Android app: <http://www.grammaticalframework.org/demos/translation.html>

The talk is based on joint work with Krasimir Angelov, Inari Listenmaa, Prasanth Kolachina, Ramona Enache, and Thomas Hallgren and partly funded by Swedish Research Council under grant nr. 2012-5746 (Reliable Multilingual Digital Communication).