Confident MT: Estimating Translation Quality for Improved Statistical Machine Translation

Irish Research Council for Science, Engineering and Technology (IRCSET)
Enterprise Partnership Scheme
http://nclt.computing.dcu.ie/mt/confidentmt.html

List of partners	
DCU	Dublin City University, Ireland (coordinator)
Symantec	Symantec, Ireland
IRCSET Investing in People and Ideas	Irish Research Council for Science, Engineering and Technology, Ireland

Project duration: November 2011 — November 2014

Summary

The aim of the Confident MT project is to develop Confidence Estimation (CE) methods to measure the reliability of Machine Translation (MT) output in the context of User-Generated Content (UGC). For localization purposes, a software company such as Symantec needs to deliver helpful content to its customers in their native languages. However, MT evaluation via automatic metrics is only possible when a reference translation is available. In the more realistic setting where no such reference is available, reliable techniques for estimating the quality of translation system output are needed. The CE methods will be applied across a range of MT systems (such as Rule-Based, Example-Based, Phrase-Based SMT and Syntax-Enhanced SMT) and the results will be used to inform the optimal combination of MT systems.

As more and more customers move away from traditional call centres and corporate websites in favour of self-service via dedicated discussion forums, there is a growing need for machine translation of UGC. Because this kind of content is an unedited mix of writing styles containing spelling mistakes, abbreviations and non-standard punctuation, it poses a particular challenge for Natural Language Processing (NLP) tools that have been trained on well-formed text.

We consider the following steps for the Confident MT project:
- Represent source and MT output text with both system-dependent and independent features.
- Adapt NLP tools (part-of-speech taggers, syntactic parsers, etc.) to UGC.
- Use particular feature classes to learn various confidence scoring models.
- Produce a confidence score to estimate machine-translated text quality.
- Combine MT systems based on CE scores.
- Examine CE scores correlation with automated and human scores.