

Automatic Improvement of Machine Translation Systems

Ariadna Font Llitjós

July 2007

Abstract

Achieving high translation quality remains the most daunting challenge Machine Translation (MT) systems currently face. Researchers have explored a variety of methods for including translator feedback in the MT loop. However, most MT systems have failed to incorporate post-editing efforts beyond the addition of corrected translations to the parallel training data for Example-Based and Statistical systems or to a translation memory database. This thesis describes a novel approach that utilizes post-editing information to automatically improve the underlying rules and lexical entries of a Transfer-Based MT system. This process can be divided into two main steps. First, an online translation correction tool allows for easy error diagnosis and implicit error categorization. Then, an Automatic Rule Refiner performs error remediation by tracing errors back to the problematic rules and lexical entries and executing repairs that are mostly lexical and morpho-syntactic in nature (such as word-order, missing constituents or incorrect agreement in transfer rules). This approach directly improves the intelligibility of corrected MT output and, more significantly, it generalizes over unseen data, providing improved MT output for similar sentences that have not been corrected.

Experimental results on an English-Spanish MT system show that automatic rule refinements triggered by bilingual speaker corrections successfully translate unseen data that was incorrectly translated by the original, unrefined grammar. Improvements on translation quality over a baseline, as measured by standard automatic evaluation metrics, are statistically significant on a paired two-tailed t-test ($p = 0.0051$).

One practical application of this research is extending and refining relatively small translation grammars for resource-poor languages, such as Mapudungun and Quechua, into a major language, such as English or Spanish. Initial experimental results on a Spanish-Mapudungun MT system show that rule refinement operations generalize well to a different language pair and are able to correct errors in the grammar and the lexicon.