Thesis Defense

Active Learning and Crowdsourcing for Machine Translation in Low Resource Scenarios

Candidate: Vamshi Ambati

Committee:

Jaime Carbonell (Chair)

Stephan Vogel (co-chair)

Raj Reddy

Aniket Kittur

Kevin Knight, (USC)

Abstract

Corpus based approaches to automatic translation such as Example Based and Statistical Machine Translation systems use large amounts of parallel data created by humans to train mathematical models for automatic language translation. Large scale parallel data generation for new language pairs requires intensive human effort and availability of fluent bilinguals or expert translators. Therefore it becomes immensely difficult and expensive to provide state-of-the-art Machine Translation (MT) systems for rare languages. In this thesis, we explore active learning to reduce costs and make best use of human resources for building low-resource MT systems. Active learning approaches help us identify sentences, which if translated have the potential to provide maximal improvement to an existing system. We then apply active learning to other relevant tasks in MT such as word alignment, classifying monolingual text by topic, extracting comparable corpora from the web. In all these tasks we reduce annotated data required by the underlying supervised learning models. We also extend the traditional active learning approach of optimizing selection for a single annotation to handle cases of multiple-type annotations and show further reduction of costs in building low-resource MT systems. Finally, as part of this thesis, we have implemented a new framework - Active Crowd Translation (ACT), a cost-sensitive active learning setup for building MT systems for low-resource language pairs. Our framework will provide a suitable platform for involving disparately spread out human translators around the world, in a timely and sparingly fashion for rapid building of translation systems. We first explore the ACT paradigm with expert translators and then generalize to full-scale crowdsourcing with non-expert bilingual speakers. In case of Machine Translation, although crowdsourcing services like Amazon's Mechanical Turk have opened doors to tap human potential, they do not guarantee translation expertise nor extended availability of translators. We address several challenges in eliciting quality translations from an unvetted crowd of bilingual speakers.