Crowdsourcing Translation

Chris Callison-Burch
University of Pennsylvania
ccb@cis.upenn.edu

Modern approaches to machine translation are data-driven. Statistical translation models are trained using parallel text, which consist of sentences in one language paired with their translation into another language. One advantage of statistical translation models is that they are language independent, meaning that they can be applied to any language that we have training data for. Unfortunately, most of the world's languages do not have sufficient amounts of training data to achieve reasonable translation quality.

In this talk, I will detail my experiments using Amazon Mechanical Turk to create crowd-sourced translations for ``low resource'' languages that we do not have training data for. I will discuss the following topics:

- Quality control: Can non-expert translators produce translations approaching the level of professional translators?
- Cost: How much do crowdsourced translations cost compared to professional translations?
- Impact of quality on training: When training a statistical model, what is the appropriate trade-off between small amounts of high quality data v. larger amounts of lower quality data?
- Languages: Which low resource languages is it possible to translate on Mechanical Turk? What volumes of data can we collect, and how fast?
- Implications: What implications does this have for national defense, disaster response, computational linguistics research, and companies like Google?

Bio

Chris Callison-Burch is an assistant professor in the Computer and Information Science Department at the University of Pennsylvania. Before joining Penn, he was a research faculty member for 6 years at the Center for Language and Speech Processing at Johns Hopkins University. He was the Chair of the Executive Board of the North American chapter of the Association for Computational Linguistics (NAACL) from 2011-2013, and he has served on the editorial boards of the journals Transactions of the ACL (TACL) and Computational Linguistics. He has more than 80 publications, which have been cited more than 5000 times. He is a Sloan Research Fellow, and he has received faculty research awards from Google, Microsoft and Facebook in addition to funding from DARPA and the NSF.