Twitter crowd translation

Ondřej Bojar
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic

Eduard Šubert
Faculty of Nuclear Sciences and Physical Engineering
Czech Technical University in Prague

June 17, 2014

1 Introduction

The aim of our project is to develop online application for creation and maintenance of corpora for machine translation without the need of hiring translators while using the power of crowd instead. The application is designed to be able to work with any source phrases; however we specialize for translation of tweets from popular social network Twitter.

2 Motivation

There are many successful crowd driven projects online; most notably Wikipedia or its sister project Wiktionary. Second example is much more important for our project since it shows that users all over the world are eager to translate in various languages. (As of June 17, 2014 site claims to have 3,766,260 entries with English definitions from over 1400 languages.) This approach could also improve role of machine translation in daily communication since it is designed to let users translate this type of phrases.

3 Our Proposal

In general there are three steps in work cycle of such application. Addition of new phrases their translation and finally their evaluation. Our solution to first step is to use social network Twitter. Application periodically scans the network for hashtag #tctrq and adds content of such tweets to database. Each tweet is required to contain another hashtag with the target language of translation. Twitter uses its own system to determine language of each tweet and our application uses this information.

Second step the translation takes place immediately thereafter. Each of registered translators capable of translation between source and target language is notified via e-mail and submits translation as a reply e-mail. These replies are periodically collected and added to database. At this point machine translation from Moses is added to compete with human translators.

Third step of work cycle the evaluation is the only step that requires user to come to our website and use simple interface to vote between two translations. Voting is of course blinded.

After gaining high enough score the translation is posted back to Twitter as a response to request and thus completing the cycle.

4 Our solution

The application is developed with the CakePHP framework. For all e-mail communication we use associated Gmail account through IMAP protocol. The Twitter integration is done with Simple PHP Wrapper for Twitter through REST API.

References

Moses - http://www.statmt.org/moses/

Twitter - http://twitter.com/

CakePHP - http://cakephp.org/