Twitter Crowd Translation – Design and Objectives

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

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

June 20, 2014

1 Introduction

This paper presents Twitter Crowd Translation (TCT), our project aimed at development of an online infrastructure serving two purposes: (1) providing online translation to social media and (2) gathering relevant training data to support machine translation of such content. We focus on Twitter and the open-source machine translation toolkit Moses. Our project heavily relies on unpaid voluntary work.

In Section 2, we provide the motivation for both goals of our work. Section 3 describes the overall design of our tool in terms of "social engineering" and Section 4 complements it by the technical aspects.

2 Motivation

Social networks have gained tremendous popularity and have successfully replaced many established means of communication. While geographical location of the users has little to no impact on communication, the obstacle of *languages used* remains.

For stable and long-lasting content, the problem is less severe: services such as the Wikipedia have shown that volunteers are able to provide translations into many languages. Machine translation is easy to train on such content and delivers moderately good results.

On the other hand, social networks are used in a streaming fashion, Twitter being the most prominent example. Anybody can contribute message, which is forwarded to a number of followers. These, in turn, are flooded with messages from sources they select. Given the constant flow of new information, nobody looks back at older messages.

Providing translation to "streaming networks" is much more challenging. The input is much noisier, significantly reducing MT output quality, and the community is less interested in providing manual translations.

The social motivation of our project is to break the language barrier for streaming social networks. The technological motivation is to advance MT quality by collecting more and better-fit data. What Wikipedia and on-line MT services manage for stable content, we would like to achieve for streaming networks and casual, unedited content.

3 Design of TCT

We see two main reasons for people to contribute to community translation of Wikipedia and other projects: sharing the information ("What is useful for me in my language may be useful for others."), and self-promotion ("I will gain good reputation by contributing well received translations."). We designed our project in accordance to these findings.

TCT should be as thin layer as possible, to cause minimal disruption. The majority of users stay within their platform – Twitter in this case.

To better explain the processes of TCT, we assign users roles: **Author**, **Selector**, **Translator**, **Judge** and **Recipient**.

Figure 1 summarizes the workflow: a tweet in a foreign language is posted by **Author** and observed by a **Selector**. The **Selector** does not fully understand the message and submits it for translation to the language of his choice. Our TCT server collects this request and forwards it to human and machine **Translators**. Translations are collected and **Judges** evaluate their quality, high-confidence machine translation might bypass this step. The best translation is tweeted to **Selector** and other **Recipients** by our server. The same user can takes several roles in the process.

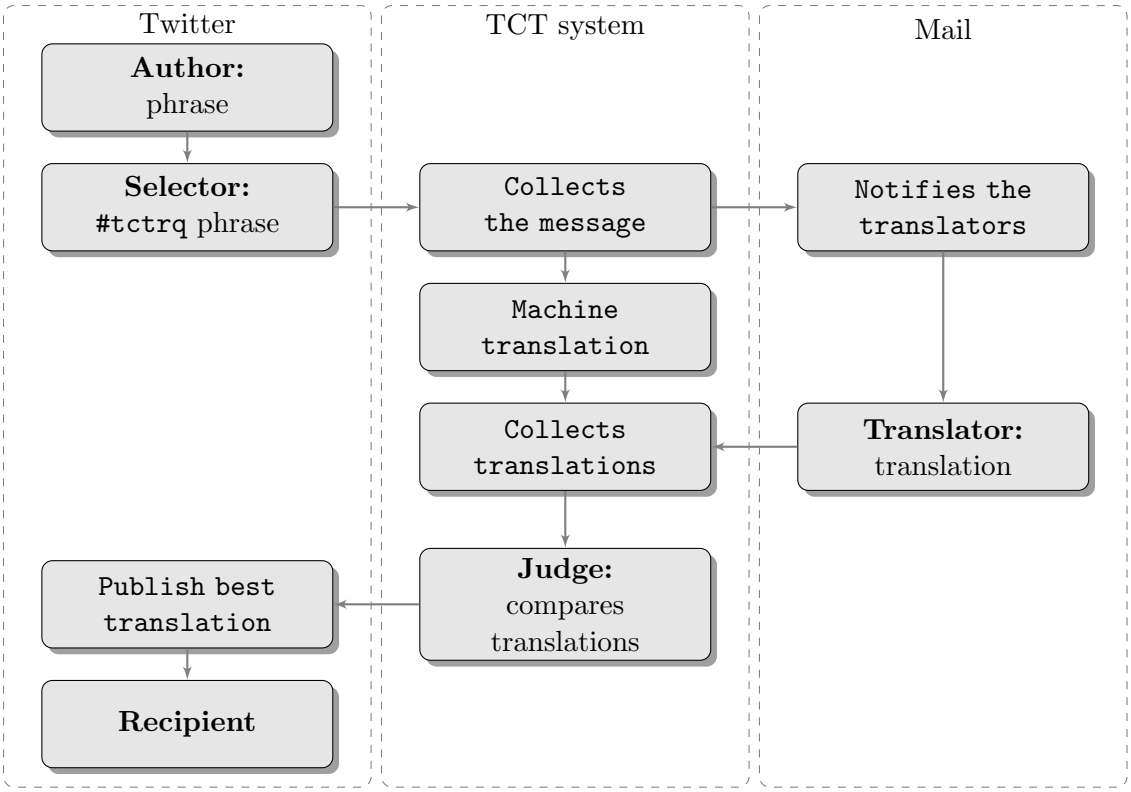

Figure 1: Twitter Crowd Translation in a nutshell.

We think that each of the user groups profits from using TCT. **Author** gains bigger audience. **Selector** achieves full understanding of tweet. **Translator** and **Judge** practice their language skills and **Translator** is placed in TCT hall of fame. Finally **Recipient** gains more of understandable content.

4 Technical Aspects of TCT

To remain in the Twitter platform, **Selector** submits messages as tweets marked with hashtag `#tctrq` and TCT uses Twitter REST API to search twitter feed for such tweets.

Once tweets are collected, **Translators** are notified via e-mail to which they respond with translations.

Judges are required to contribute via TCT website and evaluate the quality of translations by blind one-to-one comparison.

An interesting feature is password-less registration. Translators are the only group required to register but their interaction is strictly e-mail based, all necessary settings are accessed by expiring links sent via e-mail on request.

References

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

Twitter - http://twitter.com/