MT crowdsource at Yandex

Irina Galinskaya galinskaya@yandex-team.ru
Farhat Aminov aminov@yandex-team.ru
Yandex, LLC

1. Abstract

Yandex.Translate is a popular online service with a various translation scenarios and a great opportunity for crowdsourcing. Some crowdsource systems invent artificial tasks and enforce volunteers to accomplish them. We instead encourage ordinary users to run their own tasks and thus collect valuable data from natural activities in different translation scenarios. In this presentation we are going to outline the following cases:

Scenario 1: users of Yandex.Translate correct machine translation of their own text.

It is widely known, that users are not always satisfied with the results provided by machine translation systems, especially when translating texts into foreign languages. One of such cases is correspondence translation. As a result, the users unsatisfied with the translation quality copy the machine translated texts into mail clients or messengers and manually edit those translations. Only after these steps are accomplished the users would be ready to send their message to partners, clients or friends. Thus, the user edits, which represent very valuable information about incorrect translations, are carried out in third party apps and not on the translation service itself. A while back, we implemented a feature that lets our users edit translations in place, without leaving Yandex.Translate service. The feature turned out to be highly demanded and we were able to collect about 100K edits in a span of three months.

We conducted a quantitative and qualitative analysis of the gathered data and explored a possibility of using this data to improve machine translation quality.

Scenario 2: users of Yandex.Translate report errors in machine dictionaries.

In addition to the full text machine translation system, we develop machine dictionaries, both translation and monolingual. They help our users better understand complicated translations or find an appropriate replacement (e.g. synonyms) for a typed word. Evidently, information from the machine dictionaries, as well as machine translated texts, contains errors. Based on the assumption that our users would be actively engaged in improving the dictionaries, we developed a feature to easily report a wide range of dictionary related issues, such as wrong translations, capitalization errors, wrong linguistic attributes, etc. Further, we created a moderation section for professional linguists to conveniently review the reported issues. Such a two-step error reporting and verification system proved to be very effective and allowed us to significantly improve the quality of the machine dictionaries. We believe the described feature will evolve in the nearest future and will be used not only to report issues, but also to add new information to the dictionaries. To demonstrate how machine dictionaries could benefit from crowdsourcing, we analyzed the received data and accepted user contributions in various ways.

Scenario 3: Wikimedia contributors are making new articles for Russian Wikipedia by translating articles from English Wikipedia with the help of Yandex.Translate API.

Wikipedia's content translation tool has been available since the beginning of 2014 and allows contributors to translate articles into variety of languages. A year after the tool was released Yandex's machine translation technology was integrated to help contributors create new articles. Initially the feature was introduced to Russian speaking users only, but eventually it proved to be very useful and became available for numerous other languages. Recently, Wikimedia Foundation published an API that lets anyone access the post-editing data, which was produced by Wikipedia users. Based on this data, we built distributions of languages, lengths of edited fragments, edit distances between machine translated sentences and and final results published by human editors (post-edited texts). We also analyzed a possibility of using this data for training and testing of machine translation systems.