MT for Uralic Languages: Yandex Approach

Irina Galinskaya galinskaya@yandex-team.ru
Alexey Baytin baytin@yandex-team.ru
Yandex, LLC

1. Abstract

The Uralic language group is certainly interesting – it is spotted in the very distant regions of Europe and Asia. Spoken by about 25 million people, Uralic languages have spread to both sides of the Ural Mountains, reaching Balkans, Baltic region, Scandinavia, Karelia, Volga region, Western Siberia and the seaside of the Arctic Ocean.

With the exception of the three major Uralic languages (Hungarian, Finnish, Estonian), relatively small Sami and a few quite small Baltic languages (Veps, Ingrian, Livonian), all other Uralic languages are spoken on the territory of the Russian Federation. Most of them have official status in the corresponding federal regions and national autonomies. They are taught at schools, studied in universities and their usage is legally obliged in official documents. There are many enthusiasts who are trying to preserve and develop these languages. Nevertheless, it is clear that Russian is prevailing, so, in fact, most of small Uralic languages are used mostly in colloquial speech.

In terms of complexity, MT for Uralic languages is a very difficult and challenging task. First, these languages have very distinct lexicon, morphology and syntax. Second, there are many dialects, which are quite distinct as well. Third, Uralic languages have highly productive morphology which leads to a strong data sparsity in SMT. Fourth (and the worst), there are very few electronic documents available for most of these languages.

Our general approach to Uralic group was as follows. We divided its languages into three subgroups:

1) with more than 1M native speakers (Hungarian, Finnish, Estonian);
2) with 100K to 1M native speakers (Udmurt, Meadow Mari);
3) with less than 100K native speakers (Hill Mari, Karelian, Nenets).

For the major subgroup we were able to collect a sufficient amount of parallel documents from the web, and thus to build quite good baseline translation systems. Automorphology and compound splitting were used to further improve translation quality.

For Udmurt and Meadow Mari, languages from the second subgroup, we managed to crawl only a modest parallel corpora and therefore were forced to rely on a hand-crafted lexicon and morphology. The pipeline included the following steps:

1) used Bible, Wikipedia and human-made dictionaries as a main lexical base;
2) developed morphological analyzers, built lemmatized models and implemented "lemma-to-lemma" decoding;
3) post-processed lemmatized translations by synthesizing proper Russian word forms.

Lack of available documents in electronic form (and sometimes even in paper form) for languages of the third subgroup poses the question of how to build translation and language models without data. In attempts to find an answer to this question, we made a little shift to a quite specific group of small languages – magic ones. Some of them (like Elvish

dialects) have linguistic artifacts in the form of lexicon and well elaborated phonology and morphology. They are also known to be connected with Uralic languages. We considered Sindarin as an archetype of under-resourced languages and decided first to experiment with it. Some results were very promising, so we became more optimistic in upcoming efforts with MT for low-resourced Uralic languages and have successfully developed a translation for Hill Mari.