Statistical Full Text Machine Translation

Róbert HORVÁTH, Peter JURČÍK, Peter MACKO, Vladimír RUMAN, Peter SLÁDEČEK, Maroš UBREŽI, Matúš VACULA*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
adictit@googlegroups.com

This paper describes our method of statistical machine translation. There are more than 6900 different languages in the world which creates a significant language barrier and enhances the demand for an universal machine translator. Even though the history of machine translation backdates to the period of time after the world war two it still has not got fully satisfying results [1].

Today the most famous machine translator is Google Translate. In the experiment we performed on the set of random sentences Google translate obtained 74% success rate in translation. This was the best result in comparison to the Bing, WordLingo and PcTranslator. The main disadvantage is that Google API is no longer free and therefore cannot be used freely to translate large amount of text.

We have developed a method of statistical machine translation which is language independent. The most significant difference from the other statistical approaches is that we use only one sided corpus of text for the translation from one language to another. This is a huge advantage in comparison to methods used in systems like Moses which needs parallel corpus which are very difficult to find [2]. Our method preserves the advantage of statistical methods that the translated sentence is with a high probability a valid sentence of the language because it is extracted from the corpus.

Our method of text translation consists of the four main steps (see Figure 1):

1. translation of each single word,
2. generating possible candidates for translated sentence,
3. corpus lookup,
4. statistically best translation selection.

To successfully provide user with translation, our method uses simple dictionary in first step to translate each single word except for name entities or numeric expressions. Those text entities are ignored to maximize success rate of corpus lookup process because they can be replaced without text changing its meaning. In the next step all word translations are used to generate translation candidates which are searched in the corpus of target language. Speed of this process is crucial therefore the method picks up sentences for corpus lookup by their probability to become a correct translation. For this purpose the word term frequency is used as an indicator so that the sentences

* Master/Doctoral degree study programme in field: Software Engineering/Information Systems
Supervisor: Ing. Dušan Zeleník, Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies STU in Bratislava

IIT.SRC 2012, Bratislava, April 25, 2012, pp. 1–2.

with a higher value of combined probability have higher priority. In case that exact translation is not found, statistically best match is offered.

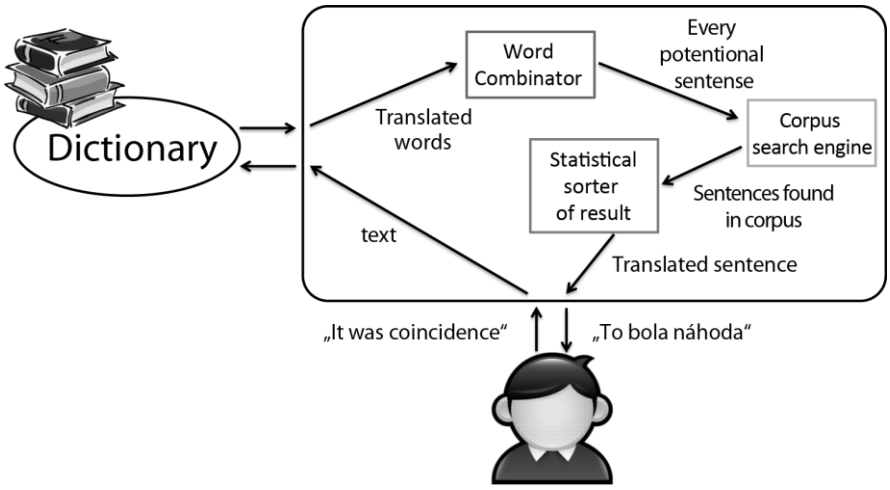

Figure 1. Process of text translation.

The most significant advantage is of our approach already mentioned language independence. The method can be extended to translate even programming or machine languages.

Disadvantage of proposed method is the time complexity which rises exponentially with the rising number of words in sentence. We are compensating this negative aspect using heuristics and distributed computation on Hadoop framework. Although the complexity of translating sentence grows exponentially with the number of words it cannot grow infinitely because every sentence has a finite word count. We are able to assess the time needed to translate the sentence with mean number of words and assume this to be constant. The complexity of text translation then grows linear depending on sentence count.

To create a solution that can reach the level of existing translators our method needs to be able to translate sentences with high success rate. We propose experiment in which we compare success rate of the most used online translators to our solution using set of pre-chosen sentences. To perform this experiment we need to enlarge our corpus.

In this paper we presented existing approaches to text translation and found that existing methods are able to translate text at maximum of 75% accuracy. We propose our statistical method which needs only one sided corpus and simple dictionary to translate from one language to another. It connects word translation, sentence generation and translation evaluation using statistical methods.

References

[1] Adam Lopez. Statistical machine translation. In ACM Computing Surveys, 40(3):1–49, ACM New York, (2008).

[2] Philipp Koehn, Hieu Hoang, et al. Moses: open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07). Association for Computational Linguistics, Stroudsburg, PA, USA, 177-180, (2007).