DANIELIUS ALGIRDAS RALYS

MACHINE TRANSLATION FOR LITHUANIAN LANGUAGE

Summary

The paper presents a historical overview as well as current state of the art of the machine translation. Translation from another language is always a certain intellectual challenge. In 1949 Warren Weaver suggested using computers to translate texts. The term "machine translation" (MT) appears. Machine translation has been rapidly developing during the first decades in order to gain a strategic advantage in the Cold War. Most popular translated languages were Russian and English. Word-to-word translation and large bilingual dictionaries covering more than 170,000 words prevailed.

In the 1950s and 1960s, machine translation systems appeared which could be called rule-based. They are based on the assumption that a language can be described using a set of rules (including grammatical). It was a rather optimistic period – it was expected to create a perfect machine translation in a few years. But a computer hardly "understands" grammar. Highly inflected languages require tens of thousands thoroughly hand-tuned and mutually consistent rules. Nobody has done this properly yet. The most advanced systems were the SYSTRAN MT system, launched in the European Commission and the Russian PROMT translation system. Subsequently, the rule-based MT progress slowed down. The EUROTRA project (1982-1992), by some estimates costing more than 50,000,000 ECU, fails – even hundreds of recruited specialists failed to create a functioning MT system. This marked a serious rule-based MT crisis. Development simply stalled for many forthcoming years.

The question was raised: if we cannot write so many rules, can it be translated without grammar at all? In 1990 a new breakthrough emerges – the research team at the *IBM Thomas J. Watson Research Center* formulates the basics of statistical machine translation. The translation process has been considered as a transmission of a certain message over a noisy channel. Decoding then has been performed on the basis of the Bayesian theorem. Translation is based on text corpora, especially on large parallel bilingual text corpora. There was a rapid improvement of statistical MT. The *EuroMatrix* project, supported by the European Commission, has created a universal open source machine translation software package MOSES, based on industry-level MT systems. Good results have been obtained – it turns out that you can translate without any dictionary or grammar! This method has greatly facilitated the translation of highly inflected languages too.

The achievements of machine translation today are effectively applied to the Lithuanian language as well. During 2005-2007 Vytautas Magnus University has carried out an EU-funded

project "Internet Information Translator". The result was a public online translation service from English to Lithuanian. (http://vertimas.vdu.lt/twsas/). The rule-based translation engine was provided by the Russian company PROMT, while other linguistic resources were prepared in Lithuania. The overall quality of text translation in BLEU metrics (in percent) is about 10. In practice, this means that only every third sentence can be adequately understood. This translation tool still has considerable potential for improving, for example by expanding phrase dictionary. Since September 25, 2008 *Google Translate* also supports Lithuanian. According to the results of the tests (2014), the BLEU translation quality was estimated to be around 17.

2012-2014 Vilnius University implemented the EU-funded project "Creation of English-Lithuanian-English and French-Lithuanian-French machine translation system based on statistical methods". The result was a public online translation service (https://www.versti.eu/). According to the tests carried out in 2014, the BLEU estimates of the translation quality exceeded more than twice the rule-based translation results and were practically equivalent to the Google translation system. When translating the documents of certain domain (such as law) the achieved BLEU score is roughly twice as high as translating general texts and far exceeds Google's results (19-09-2014). However, even the best machine translations often require human intervention and final editing in order to get the perfect translation. So, can machines ultimately translate fine?

The last few years promise new breakthroughs using a neural machine translation. Neural networks themselves construct transformation rules. It is likely that in the near future the neural MT systems will translate better than an average translator. In 2018 Vilnius University is preparing to launch the EU-funded project of a new generation neural machine translation of English, Lithuanian, Polish, French, Russian and German.

Thus, the latest achievements in machine translation apply to the Lithuanian language as well.

KEYWORDS: Lithuanian, machine translation, computational linguistics, history, neural networks, artificial intelligence.

DANIELIUS ALGIRDAS RALYS

Vilniaus universiteto Taikomųjų mokslų institutas

M. K. Čiurlionio g. 29, 03100 Vilnius

danielius.ralys@gmail.com