A Possible Solution to the Problem of Machine Translation of Verb Forms from Bulgarian to English

Todor Lazarov
Department of Computational linguistics
Institute for Bulgarian Language
Bulgarian Academy of Science

Why a problem?

- Translating verb forms is very difficult even for human translation – even though the verb systems of both English and Bulgarian share numerous common characteristics, they differ in the manner in which they express the relations between events and points on the temporal axis, the action denoted by the verb and the information about these events. Nevertheless, as we speak about the opportunities of machine translation, both languages are resource rich, which makes theoretical and practical researches about different aspects of them reliable and the gathered data – practical for the purposes of natural language processing and machine translation.

Transfer-based rules

- As it has been pointed out before, the characteristics of the grammaticalized information in Bulgarian and English verb forms share numerous similarities. That is why we have similar grammatical meaning in most of the verb forms and thus we can construct sufficient transfer-based rules for translation. We have to point out again that most of the grammaticalized information is lost or its semantic has to be converted between different grammatical categories during the semantic transfer between both languages, which makes the choice of the proper transfer-based rules inaccurate:

- *V:pres_2_s* →*V:pres_0*
- *пишеш*→*write*
-
- *V:part_aor_f_s + AuxV:pres_1_s* →*AuxV:pres_0+V:part_past*

  писала съм→*have written*

Statistical language modeling

- The goal of statistical language modeling is to build a statistical language model that can estimate the distribution of natural language as accurate as possible. A statistical language model is a probability distribution *P(s)* over strings *S* that attempts to reflect how frequently a string *S* occurs as a sentence. By expressing various language usages and deviations in terms of simple parameters in a statistical model, it can provide an easy way to deal with complex natural language phenomena. For the purposes of statistical language modeling we need to know whether a given string of words is the right string of words in given language – we need to know what the probability of the string is.

A possible solution ?

- We already pointed that transfer-based rules can provide information about the exact semantic and lexical transfer between the two languages of interest for us, nevertheless, in the case of translating from Bulgarian to English they cannot be prescribed with 100% certainty due to the huge amount of grammatical information that is lost during the process of translation, thus we need to construct a statistical language model of the transfer-based rules on their own. After that we could generate translation model, which is going to rely on the transfer-based rules. By analyzing data from enough monolingual corpora, we will be able to construct statistical language models for Bulgarian and English. After that, we will need to gather data from parallel corpora and apply it to the language models we have constructed in order to construct a statistical language model of the verbs.

- By analyzing the statistical language models of the verbs in both languages, we will be able not only to achieve better quality of machine translation between Bulgarian and English, but also to answer many theoretical questions about the grammatical dependencies between these two languages.