***R. Garabík***

# A SIMPLE RUSSIAN-SLOVAK MACHINE TRANSLATION SYSTEM

## 1. Introduction

Modern approaches to machine translation tend to favour statistical methods, usually exploiting huge parallel corpora. This is understandable because a rule-based system requires a skillful linguist to write the rules and depends on a fairly precise and detailed analysis of the source text in order to have enough input data for the rules to match. On the other hand, statistical methods require «only» sufficient amount of training data (and suitable algorithms) – yet, some level of nontrivial analysis of the input text is still required. Usually, the best results are achieved using combined methods.

Nevertheless, parallel corpora exist only for «big» languages – while frequently a pair with English language exists, for smaller languages there are often not any parallel corpora available. What is available for many languages, however, is a morphology analyzer or a word form generator, being a basic requirement for almost any NLP related research. Since closely related languages possess common syntactic, morphological and lexical features, required machine translation transfer is greatly reduced. Commonly, for very close languages, only a dictionary translation of lemmas and one-to-one mapping between morphological categories is enough to get a working translation system[1].

[1] *Hajič J., Hric J., Kuboň V.* Machine translation of very close languages // Proceedings of the Sixth Conference on Applied Natural Language Processing. Seattle, Washington. Morg an Kaufmann Publishers Inc., San Francisco, 2000. P. 7–12.

## 2. Dictionary and morphology mappings

We used a bilingual dictionary containing translations of about 73 000 Russian words with a level of homonymy 2,98 (i.e. one Russian entry was translated on average by 2,98 different Slovak translations). As the data came from a general-purpose dictionary, it exhibits several features adverse to the machine translation purposes. First, the dictionary tries to cover the *meaning* of words as thoroughly as possible, translating one Russian word with several Slovak ones, provided the semantic meanings overlap (even if considering only some obscure or rare usage). This has the unfortunate consequence of unnecessary increased translation ambiguity and subsequent lower translation quality. Second, the dictionary contains many typos and evident mistakes. We corrected some of the most obvious and frequent ones but many still remain.

For each word in the dictionary, we generated all the possible inflected forms for both Russian[1] and Slovak[2], and paired the forms according to morphological categories. For most of the part of speech categories, the mapping was straightforward – gender and number for nouns, indicative, infinitive and imperative forms, person and number for verbs, number and degree for adjectives, degree for adverbs. There are just several prominent exceptions:

– In Russian, the genitive case is used after negated verbs, while in Slovak it is the accusative that is used in this situation. This has been acknowledged by pairing Russian genitive with both Slovak genitive and accusative.

– Some Russian prepositions govern different grammatical cases. As an example we can take the Russian preposition *после* (*=after*) governing the genitive, while corresponding Slovak *po*

[1] **http://sourceforge.net/projects/phpmorphy/**

[2] *Garabík R.* Slovak morphology analyzer based on Levenshtein edit operations // Proceedings of the WIKT'06 Conference. Bratislava, Slovakia, 2006. P. 2–5.

governs the locative. We have expanded the morphology mappings by explicitly generating the correct pairs with the preposition included for all the parts of speech with the case category, e.g. *после весны*↔po jari (=*after spring*)[1].

– Adjectives have to agree in gender with nouns, however, translated nouns often differ in gender. Therefore, we have paired adjective word forms regardless of their gender.

– The same can be said about the only verb form exhibiting gender distinction, the so called *L*-participle (used to construct the past tense and conditional) in singular, which has to agree in gender with the sentence subject.

Together, about 50 relatively simple and straightforward rules have been needed to describe the complete pairing. After morphology expansion, the number of Russian entries has expanded to ~300 k words and the level of homonymy reached 11,7.

### 3. Language model

We have used a simple second order Markov chain language model, i.e. the probability of an $n^{\text{th}}$ word ($x_n$) in a sentence is given as a transition probability depending on two previous words, and we are maximizing the probability of a sequence of words in the sentence of the length $N$:

$$translation = \underset{x_1 \ldots x_n}{\arg\max} \prod_{i=1}^{N} P(x_i \mid x_{i-1}, x_{i-2})$$

$x_0$ and $x_{-1}$ are special pseudotokens denoting a sentence beginning. The most probable sequence is found by a slightly modified Viterbi algorithm – we do not construct a trellis, but calculate the

[1] The Slovak *po* can also govern the genitive, but with different meaning (to send *after* someone), which corresponds to Russian preposition *по* (governing the genitive, too). However, this is taken care of automatically by pairing genitive cases by default.

transition probabilities at each step dynamically, the set of possible next states being the set of possible translations for a given source word.

The transition probabilities are obtained from a trigram distribution of a target language corpus with linear interpolation smoothing:

$$P(x_n \mid x_{n-1}, x_{n-2}) = \lambda_3 f(x_n \mid x_{n-1}, x_{n-2}) + \lambda_2 f(x_n \mid x_{n-1}) + \lambda_1 f(x_n)$$

where $\lambda_i$ are weight for tri- bi- and unigram probabilities such that $\sum_i \lambda_i = 1$ and can be obtained from the training data by deleted interpolation technique.

### 4. Evaluation

For the evaluation, we have chosen the METEOR benchmark[1], an evaluation metric that should give high levels of correlation with human judgment. The evaluation was intentionally performed using exact match on surface word forms, not on lemmas, in order to be able to estimate improvements in morphology of translated texts. In the absence of Slovak language parameters, we have used the Czech ones, relying on the validity of the results due to language similarity[2]. The evaluation has been performed on a fixed set of 1000 randomly chosen perfectly aligned sentences obtained from the parallel Russian-Slovak corpus[3].

[1] *Lavie A., Agarwal A.* METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. Proceedings of the Second Workshop on Statistical Machine Translation at the 45th Meeting of the Association for Computational Linguistics (ACL–2007), Prague, Czech Republic, 2007. P. 228–231.

[2] Measurements of relative improvements should be valid anyway.

[3] *Garabík R., Захаров В.П.* Параллельный русско-словацкий корпус. Proceedings of the Conference «Corpus Linguistics–2006». P. 81–87. St. Petersburg University Press, St. Petersburg, 2006.

## 5. Transliteration

Vast majority of Russian proper nouns, when used in a Slovak language text, is usually transliterated more or less according to the scheme outlined in the Rules of Slovak Orthography[1] (we denote this transliteration scheme as *PSP*). Since only a few of the most common names are present in the bilingual dictionary and recognized by the morphology analyzers, by introducing transliteration of unknown words into the translation process we hope to improve the translation efficiency, considering the fact that thanks to inherent similarity between Russian and Slovak lexicon, the transliteration will transcribe correctly at least some percentage of unknown words (in addition to proper nouns). Of course, the transliteration will probably help only as far as basic word forms are concerned, because of the difference in morphological suffixes. Indeed, we gained a noticeable improvement in the translation score, as seen in *Table 1*. We can achieve yet another improvement by taking into account regularity of orthography differences between Russian and Slovak and designing our transcription rules in order to maximize similarities between transliterated Russian words and their Slovak counterparts (e.g. by transliterating word final *-ся* as a separate word *sa* instead of PSP-compliant *-sia* we hope to recover some percentage of reflexive verbs). Overall, the transliteration consists of about 150 simple string substitution rules. As the *Table 1* shows, this leads to further improvement of the translation (we denote this improved transliteration as *genetic*). We expected that by combining these two transliterations we can further improve the translation. By providing both possibilities and letting the disambiguation process choose the most probable one we hoped to select the PSP form for proper names and the genetic form for some of the unknown words (if present in the *n*-gram data from the target language corpus). However, this turned out not to be the case, and the

[1] Pravidlá slovenského pravopisu. Ed. M. Považaj. Bratislava. Veda, 2000.

combined transliteration gives in fact slightly lower score. The reason might be caused by increased translation ambiguity and inconsistency, not counterbalanced by better vocabulary coverage.

*Table 1.* Different transliteration possibilities. The first three entries contain only raw transliteration, without any translation at all, and are included for reference purposes. We have also included an evaluation of the translations of a tiny (10 sentences) subset of the reference test sample, made independently by two human translators.

| **Type** | **METEOR score** |
|---|---|
| PSP transliteration only | 0,1392 |
| Genetic transliteration only | 0,1523 |
| Combined transliteration only | 0,1490 |
| Translation only | 0,2646 |
| Translation + PSP transliteration | 0,3243 |
| Translation + genetic transliteration | 0,3322 |
| Translation + combined transliteration | 0,3279 |
| Human 1 | 0,5280 |
| Human 2 | 0,5103 |

## 6. Simulating different dictionary properties

When considering qualities of any system, it is important to find out the robustness with regard to input conditions. We tried to simulate the influence of different bilingual dictionary sizes on the overall translation quality by selecting different amount of unique random entries from the dictionary. The results are summarized in *Table 2* and *Figure 1*. It is clear that saturation level has not been reached and the translation can be still improved by increasing the dictionary size.

*Table 2.* Score as a dependency on dictionary size

| **Size [%]** | 0 | 20 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|
| **score** | 0,1523 | 0,1876 | 0,2166 | 0,2503 | 0,3015 | 0,3322 |

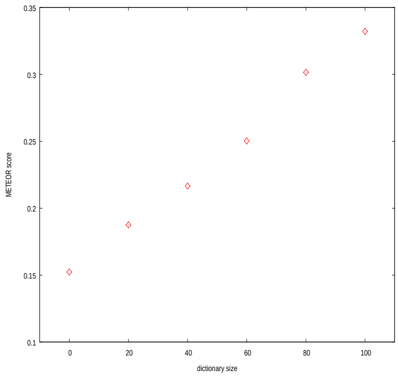

*Fig. 1*. Score as a dependency on dictionary size

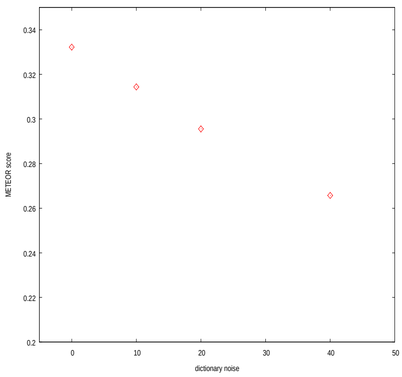

*Fig. 2*. Score as a dependency on dictionary noise

We also tried to estimate the effect of misplaced dictionary entries. For this purpose, we randomly chose a subset of dictionary entries and randomly shuffled Russian and Slovak lemmas in this subset. The results are displayed in *Table 3* and *Figure 2*. The presence of «noise» in the dictionary has predictable results, the (relatively) low sensitivity of the overall score to the noise can be explained by the presence of the morphology pairing algorithm, which discards all the word pairs with disagreeing part of speech categories.

*Table 3*. Score as a dependency on dictionary noise

| **noise [%]** | 0 | 10 | 20 | 40 |
|---|---|---|---|---|
| **Score** | 0,3322 | 0,3144 | 0,2956 | 0,2657 |

## 7. Incompatible grammar features

One of the most distinguishing features of Russian syntax is the zero copula in the present tense. The old Slavic conjugated copula is not used anymore (only to create an archaic effect), 3rd person singular *есть* is used instead for all persons and numbers, but is limited to existential usage and emphasis. Often, the zero copula is represented in orthography by a dash. In Slovak, copula is a full featured compulsory verb, omission of which is acceptable only as an ellipsis. We tried to reconcile these differences by introducing Slovak copula as one of the possible translations of Russian personal pronouns (the other being the sole pronoun), relying on the underlaying language model to pick up the most probable word sequence, with very good results. In all the other usages, the copula remains absent from the translated sentences, but the resulting syntactical sentence structure is perceived as an ellipsis by native speakers (reinforced by the dash, if present) and is not distracting in a major way.

In Russian, the possession is not expressed with a verb corresponding to an English «*to have*». Even if the equivalent word exists (*иметь*), its usage to describe possession is marginal at best. The standard way is to use the preposition *у*, possessor in genitive and the object possessed in nominative. Slovak uses the verb *mat'* with an accusative construction. Direct translation of the Russian construction into Slovak is impossible without at least some level of syntactic analysis, which is out of the scope of this translation system. Fortunately, the Russian-like construction, although unusual, is grammatical in Slovak (if we broaden the definition of grammaticality a bit). We decided, therefore, to ignore this discrepancy relying on the intelligibility of the final word for word translation.

## 8. Conclusion

Despite its simplicity, presented system achieves a good translation quality. Inherent syntactical differences between Russian and Slovak do not pose major obstacles to the intelligibility of the translated text and can be alleviated by choosing appropriate morphology pairing rules. The system is viable for translation between similar languages and we foresee its application for pairs of different Slavic languages.