Statistical machine translation between related and unrelated languages*

David Kolovratník, Natalia Klyueva and Ondřej Bojar

Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Abstract. *In this paper we describe an attempt to compare how relatedness of languages can influence the performance of statistical machine translation (SMT). We apply the Moses toolkit on the Czech-English-Russian corpus UMC 0.1 in order to train two translation systems: Russian-Czech and English-Czech. The quality of the translation is evaluated on an independent test set of 1000 sentences parallel in all three languages using an automatic metric (BLEU score) as well as manual judgments. We examine whether the quality of Russian-Czech is better thanks to the relatedness of the languages and similar characteristics of word order and morphological richness. Additionally, we present and discuss the most frequent translation errors for both language pairs.*

1 Introduction

Statistical Machine Translation nowadays has become one of the easiest and cheapest paradigms of the MT systems. Researchers can now use various toolkits to experiment with different language pairs. We experiment with Moses [2], an open-source implementation of phrase-based statistical translation system.

For closely-related languages, statistical MT methods are sometimes believed to be unreasonably complicated. For example, in the project Česílko [3] – Machine Translation among Slavic languages – the main accent was put on the idea that the relatedness of the languages rather than statistics should be exploited. Česílko was initially a rule-based system, based on the direct word-for-word translation (for very closely related Czech and Slovak) and engaging a few syntactic transfer rules in case less related languages are concerned (Czech and Polish or Czech and Lithuanian).

In our experiments we try to compare if the relatedness has a positive effect when using phrase-based statistical models.

Our main hypothesis was that we should obtain better results in Russian-to-Czech translation than in English-to-Czech. We used the Moses toolkit in order to carry out the experiments and evaluation. Additionally, we applied factored models on the tagged version of the corpus and compared the outputs.

The paper is structured as follows. Section 2 and Section 3 provide a description of the data we used during the experiment and our tokenization and tagging tools. In Section 4 and Section 5 we briefly summarize the Moses toolkit and present our experiments with MT between English/Russian and Czech. In Section 6 we evaluate our MT output using an automatic and a few manual evaluation metrics. Finally, the paper is concluded by a discussion and plans of future work.

2 Data

Phrase-based SMT systems need huge amount of parallel data in order to extract dictionaries of phrases and their translations, so called phrase tables. The most reliable source of parallel data are books and their translations into different languages, still it seems to be very laborious to collect a big corpus based on books. Web pages can serve as a good and significantly cheaper source for parallel texts, although usually less reliable. Moreover, while for the wide-spread languages we can easily find them, for minority languages parallel texts may not be available on the web in sufficient quantities.

We carried our experiments using the Czech-English-Russian (cs-en-ru) corpus UMC 0.1 [1] with automatic pairwise sentence alignment containing texts from Project Syndicate[1]. Although we could have used additional data to train the translation model for Czech and English, we need English-Czech and Russian-Czech corpus to be comparable. Table 1 provides statistics of the data we used in our experiments.

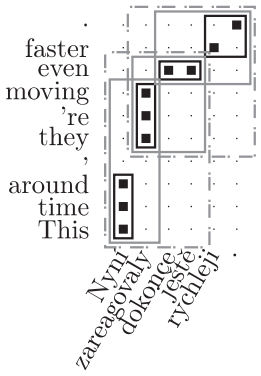
We had to collect the held-out and test set sentences ourselves for two reasons: first, we needed the sentences to be tri-parallel, that is parallel across the three languages, and second to be sure they do not overlap with the training data set. We also used Project Syndicate but extracted the test sets only from

* This work was supported by the Czech Science Foundation under the contract no. 201/09/H057, Grant Agency of Charles University under the contract no. 100008/2008, and the grants FP7-ICT-2007-3-231720 (EuroMatrix Plus), GAAV CR 1ET201120505 and MSM 0021620838.

[1] `http://www.project-syndicate.org/`

Cz: *prostě/prostě/Dg-------1A---- jsem/být/VB-S---1P-AA--- brala/brát/VpQW---XR-AA---*
Ru: *включая/включая/Sp-a президента/президент/Ncmsay мбеки/мбеки/Vmip3s-a-p*
En: *the/the/DT visionaries/visionary/NNS would/would/MD have/have/VH gotten/get/VVN nowhere/nowhere/RB*

Fig. 1. Example of a factored corpus. The sentences are not parallel.



This	=	nyní
time	=	nyní
around	=	nyní
they	=	zareagovaly
...	=	...
This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

Fig. 2. Simple phrase-based translation: Training sentences are automatically word-aligned and used to extract all phrases constistent with the word alignment (not all consistent phrases have been marked in the picture). The extracted dictionary of phrases is used in translation: the input sentence is segmented into known phrases, each phrase is translated and the output is constructed by concatenating translated phrases. Usually only little phrase-reordering is performed.

	Languages	Sentences
Language Model	cs	92,233
Translation Model	ru → cs	79,888
Translation Model	en → cs	76,588
Held-out	cs, en, ru	750
Test set	cs, en, ru	1,000

Table 1. Summary of corpus sizes.

newly published articles. The held-out and test set sentences have been added to the corpus UMC[2].

3 Data preprocessing

We used the tools developed under the UMC project, namely the trainable tokenizer for Czech, English and Russian languages. It was applied on the test and development set of data to make them consistent with training sets.

In order to train a factored model we tagged and lemmatized the UMC corpus with the help of TreeTagger [5] for English and Russian and Hajič's morphological tagger for Czech [8]. Figure 1 provides examples of the tagged and lemmatized parts of text in the format as suitable for the factored training.

4 Simple Moses

Moses[3] is a phrase based SMT system that is very much language independent since it implements a purely data driven method. In contrast to other methods of MT, phrase-based systems can perform translation directly between surface forms (thus often the name "direct translation"). The most important property of phrase-based systems is the ability to translate contiguous sequences of words (called "phrases") rather than merely single words. See Figure 2 for an illustration.

The Moses toolkit is a complex system which utilizes several other components. Let us mention at least GIZA++[4] involved in finding word alignment, the SRI Language Modeling Toolkit[5] and the built-in implementation of model optimization (Minimum Error Rate Training, MERT) on a given held-out set of sentences.

To establish a baseline, we trained translation models for direct translation from Russian to Czech (ru→cs simple) and English to Czech (en→cs simple), optimizing them on the 750 held-out sentences.

5 Moses factored

All knowledge used by Moses comes from the corpus. Moreover, direct phrase-based translation models have no generalizing capacity. Thus their performance strongly depends on whether particular words and word sequences were seen in the training sentences data. Phrase-based translation thus often faces a problem known as data sparseness, and the problem is more

[2] http://ufal.mff.cuni.cz/umc/
[3] http://www.statmt.org/moses/
[4] http://www.fjoch.com/GIZA++.html
[5] http://www.speech.sri.com/projects/srilm/

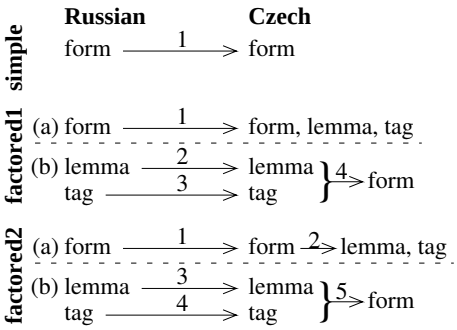


Fig. 3. Ilustration of all explored translation settings: (a) and (b) parts represent alternative decoding paths of a given factored setup.

pronounced for morphologically rich languages where all word forms have to be seen.

Factored translation [6] is an interesting extension of phrase-based models that aims i.a. to mitigate this issue. It allows us to replace an input word with a vector of features as exemplified in Figure 1 and configure the model to back-off to a more coarse-grained representation of input words if there are not enough training data. The features on the source side can also participate in translation. Features on the target side may be obtained by translation from the source side or by a generation step. The generation works with features already available on the target side and fills in the remaining ones.

The most common example of employing factored translation looks as follows. A surface word form is enriched with its base form (lemma) and morphological information (a tag for short), forming a three-compound features vector. Base forms and tags are translated independently without regard to surface forms. Then, on the basis of translated base form and tag the surface form is generated. The setup can use three language models ensuring coherence of the output sequence: one for base forms, one for tags and one for surface forms.

To summarize, there are two translation models (for base forms and for tags), one generation table to get surface form and three language models. This was the approach we first planned to exploit. Unfortunately, the setup has a subtle drawback: it does not work with input forms at all, so it applies the independent translation of base form and tag even in cases where there is enough data for direct translation. Moses allows to specify multiple decoding paths (decoding means finding the most probable translation of a given sentence according to the model), so it is possible to let compete the factored path with the direct transfer, exploiting mutual advantages of both approaches. That is the approach we used in our factored experiments.

Although in the direct translation path used as the back-off of the factored translation we are not interested in the target-side lemma and tag, we still have to supply them for the language models. We use two distinct setups for constructing the additional output factors for the direct translation: 1) translating the source form to all three target factors at once, and 2) translating the source form to target source form and using a generation step for "instant tagging" of the output to construct the target lemma and tag. We denote the combination of the main factored translation with one of the two back-off models *factored1* and *factored2*, resp. Both are ilustrated in Figure 3.

We are aware that there is relatively little possibility for an improvement with factorization in our language pairs and overall setting. For instance, let us point out that generation step for target-side factors is integrated into Moses unlike the preprocessing of input factors where external tools are used. Naturally, the generation capabilities of Moses are rather limited: it learns only from sentences supplied in training. Because we train the generation step only on the target side of the parallel sentences, we cannot expect to gain much coverage by translating lemmas and tags independently because the data will hardly ever provide the required form that should be generated from the target lemma and tag. A better approach would be to either use a larger monolingual corpus for training the generation step, or use an external morphological generator as e.g. [9]. With the current simple setting, we can expect improvement rather to come from the additional lemma- and tag-based language models that will be able to judge hypothesis coherence more robustly.

6 Evaluation

We tried to evaluate the output of our systems by several metrics: BLEU, flagging of errors and a simple hypothesis ranking (i.e. asking "which is the best output").

6.1 BLEU

BLEU score [4] is an established automatic metric used to evaluate MT systems. Thus, despite all known issues we also used it not only for completeness but also as an integral part of model optimization (see MERT in Section 4). Anyway, let us mention two major issues of the BLEU score.

BLEU, when applied to languages with free word order, cannot be reliable indeed. BLEU is based on counting occurrences of n-grams from reference translation in generated output. In many cases the

translator of reference texts will use a word order different from the source sentence, whereas the machine usually preserves the original word order whenever it is an acceptable variant. However, many n-grams do not match when words are swapped. Here are some examples of the problem from our test data:
(reference translation) *syrský postoj by dosah íránské strategie regionální destabilizace nemusel rozšiřovat, ale spíš omezovat.*
(ru→cs translation) *postoj sýrie může omezit, nikoliv rozšířit, sféru vlivu íránské strategie regionální destabilizace.*

Such shifts done by a translator lead to a lower (automatic) score while not necessarily impacting the comprehensibility of the output.

There is a similar problem with inflection. Word forms different from the reference translation are not approved by the BLEU score, so minor translation variations or errors can cause unfair loss in BLEU score. However, a partial remedy may be achieved by scoring lemmatized text:
(reference translation) *složitost hrozeb , jimž čelí izrael*
(ru→cs translation) *složitost hrozeb izraeli*
(en→cs translation) *složitostí hrozby pro izrael*

Table 2 summarizes BLEU scores obtained by our various translation setups. For English all scores are very close. In contrast, Russian is more sensitive to a method – factored translation performs slightly better than simple. Unfortunately, we were unable to compute *factored2* for Russian due to troubles with model optimization. A discussion of closeness of simple and factored results is to be found in the last paragraph of Section 5.

	BLEU score on forms		
pair	simple	factored1	factored2
en→cs	14.58±0.96	15.84±1.03	15.39±1.05
ru→cs	11.91±0.91	13.11±0.90	—

	BLEU score on lemmas		
pair	simple	factored1	factored2
en→cs	24.16±1.10	24.77±1.18	24.99±1.16
ru→cs	15.98±0.97	18.06±0.92	—

Table 2. Achieved BLEU scores in our experiments.

6.2 Flagging of errors

As shown in the previous section, the BLEU metric does not always reflect translation quality. A more reliable, though labour-intensive approach is to manually judge MT output. In one of such evaluations, inspired by [7], human annotators mark errors in MT output and classify them according to their nature. We used the following rough error classes: **Bad Punctuation**, **Unknown Word**, **Missing Word**, **Word Order**, **Incorrect Words**, with some classes further refined into several subtypes. As our annotation capabilities were limited to one person only, we present here the evaluation of the simple model (direct translation) only.

Table 3 documents that in the case of English-to-Czech translation, the most common errors concerned morphology, which matches our expectations as Czech is a inflective language and needs to express many features like case and gender, often not marked in English source. On the other hand, lots of words were not recognized in Russian-to-Czech translations. We have not been able to evaluate the factored translation according to the scheme, but a first few sentences show higher accuracy in morphological forms when factored models are used.

Error Class	en→cs	ru→cs
Disambiguation	9.3 %	8.8 %
Extra word	6.2 %	18.2 %
Word Form	49.0 %	22.0 %
Lexical Variant	5.4 %	5.7 %
Missed Auxilary	0.8 %	1.9 %
Missed Content	6.6 %	20.1 %
Word Order Long	0.8 %	0.6 %
Word Order Short	4.6 %	0.6 %
Punctuation	13.9 %	2.5 %
Unknown	3.5 %	19.5 %
Total	259 (100.0%)	159 (100.0%)

Table 3. Error types in simple moses model.

6.3 Ranking of translations

Finally, we carried out a ranking evaluation which is very similar to the human judgments in WMT Manual Evaluation[6]. For each of the translation schemes described in Section 4 and Section 5 we took 40 sentences and ranked them on the basis of the question "which translation is the best". So each MT output of the 40 test sentences translated to Czech from both languages and by all examined setups got a score from 1 (worst) to 5 (best). Table 4 summarizes the evaluation. For each translation setup, we compute the mean, median and count of how often the method got the best and the second best rank.

Almost a half of the sentences that got the highest score were factored translations from Russian into

[6] `http://www.statmt.org/wmt08/judge/`

En→Cz	simple	factored1	factored2
Median	3	3	2
Mean	2.487	3.051	2.718
Best/Second	2/8	9/6	4/6

Ru→Cz	simple	factored1	factored2
Median	4	4	—
Mean	3.436	3.923	—
Best/Second	10/12	19/9	—

Table 4. Manual ranking of MT output.

Czech, the second score was obtained by those translated using the simple model from Russian into Czech. Factored model (*factored1*) from English to Czech was the third one. This confirms our expectation that translating from a related language is easier also for phrase-based MT.

The evaluation allows us to make further conclusions. First, enriching the model with additional morphological information improves the translation quality both for related and unrelated languages. For Russian as the source, the improvement seems to be less apparent, because Russian itself marks most of the relevant morphological properties in its word forms. Second, BLEU score does not necessarily corresponds with manual judgments: while translating from Russian was better percieved by our human annotator, it obtained a lower BLEU score than translation from English[7]. We are aware that the evaluation should be repeated with more human annotators and on a larger set of sentences for a better confidence.

6.4 Observation of frequent errors

As it was shown in the previous section, there are lots of words unrecognized (not translated). This problem is not of a linguistic nature, it is caused simply by insufficient training data.

Here we will name some linguistically interpreted errors.

- Russian → Czech
  - Lost negation.
    (ru src) *без которого было невозможно создание*
    (cs ref) *bez něhož* **ne***bylo možné sestavit*
    (ru → cs) *bez něhož bylo možné vytvoření*
    Here we can observe that due to the difference in how negation is expressed in the two languages, the negative sense is translated as positive.
  - Lost reflexive particle.
    (ru src) *сумел уйти от*
    (cs ref) **se** *zdařilo vyjít z*
    (ru → cs) *podařilo odejít od*
    The mistake above—missing reflexive particle in Czech—is caused by the fact that some verbs can be reflexive in Czech and non-reflexive in Russian which is difficult for a phrase-based MT to learn because the reflexive particle is often far away from the verb in training sentences.
- English → Czech
  - Word order in possessive constructions.
    (en src) *mahmoud abbas 's palestinian authority*
    (cs ref) *palestinskou samosprávou prezidenta mahmúda abbáse*
    (en → cs) *prezidenta mahmúda abbáse palestinské samosprávy*
- Both source languages → cs
  - Bad case after a preposition.
    (cs ref) *podle indických vyšetřovatelů*
    (en src) *according to indian investigators*
    (en → cs) *podle* **indické** *řešitelů*
    (ru src) *согласно индийским экспертам*
    (ru → cs) *podle* **indickým** *experti*

7 Conclusion

We have succeeded in our goal to compare the performance of phrase-based and factored phrased-based statistical machine translation when translating between related and unrelated languages. So far we have failed in taking advantage of language relatedness explicitly in the model, but a preliminary manual ranking of system outputs confirms that translation between related languages delivers better results. This observation contradicts to the automatic MT quality score using the BLEU metric.

We are aware of the remaining data sparseness issue (there are many times more tags for Russian than for English), so while the language relatedness makes the Czech and Russian tagsets similar, many tags needed in the translation of unseen sentences are not in our training data. Also we suspect the training corpus to be better parallel for English-Czech pair than for Russian-Czech, because Czech is the direct translation of English original while Russian is the translation of English, not Czech.

Our second conclusion is that enriching SMT with morphological features improves the translation quality especially for the closely-related morphologically rich Czech and Russian.

[7] While BLEU scores are not comparable across language, they *are* comparable in our setup: we test BLEU scores on a single test set in Czech only, it is the source language that differs, not the target one.

We hope that our results will serve as a good basis for a future comparison of SMT with rule-based approach used in Česílko, which intends to include Russian-Czech translation pair soon. Our experiments are also a good start for further improvements in MT quality when translating to Czech. For instance, we plan to improve the morphological generation step by using larger target-side monolingual training data.

References

1. N. Klyueva and O. Bojar: *UMC 0.1: Czech-Russian-English multilingual corpus.* Proc. of International Conference Corpus Linguistics., Saint-Petersburg, 2008, 188–195.
2. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst: *Moses: open source Toolkit for statistical machine translation.* ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, 2007, 177–180.
3. P. Homola and V. Kuboň: *A hybrid machine translation system for typologically related languages.* Proceedings of the 21st International Florida-Artificial-Intelligence-Research-Society Conference, FLAIRS, 2008, 227–228.
4. K. Papineni, S. Roukos, T. Ward: *BLEU: a method for automatic evaluation of machine translation.* IBM Research Report RC22176(W0109-022), 2001.
5. H. Schmid: *Probabilistic part-of-speech tagging using decision trees.* Proceedings of International Conference on New Methods in Language Processing, 1994.
6. P. Koehn and H. Hoang: *Factored translation models.* Conference on Empirical Methods in Natural Language Processing (EMNLP), 2007, 868–876.
7. D. Vilar, J. Xu, L. Fernando D'Haro, and H. Ney: *Error analysis of statistical machine translation output.* LREC-2006: Fifth International Conference on Language Resources and Evaluation. Proceedings, Genoa, Italy, 22-28 May 2006, 697–702.
8. J. Hajič: *Disambiguation of rich inflection.* (Computational Morphology of Czech). Nakladatelství Karolinum, ISBN 80-246-0282-2, Prague, 2004.
9. A. de Gispert, J.B. Mariño and J.M. Crego: *Improving statistical machine translation by classifying and generalizing inflected verb forms.* Eurospeech 2005, Lisbon, Portugal, 2005, 3185–3188.