Evaluating Quality of Machine Translation from Czech to Slovak*

Ondřej Bojar, Petra Galuščáková, and Miroslav Týnovský

Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Praha 1, CZ-11800, Czech Republic
{bojar,galuscakova,tynovsky}@ufal.mff.cuni.cz

Abstract. *We focus on machine translation between closely related languages, in particular Czech and Slovak. We mention the specifics of evaluating MT quality for closely related languages. The main contribution is the test and confirmation of the old assumption that rule-based systems with shallow transfer still work better than current state-of-the-art statistical systems in this setting, unless huge amounts of data are available.*

1 Introduction

Machine translation between closely related languages has been studied long in the past (e.g. [8]). From the technical point of view, the main motivation is the easier transfer (or no transfer at all for some cases), so hopefully better output quality. Savings can be expected in setups where a single document is translated into many target languages: human translators can provide the translation to a "pivot" language and subsequent MT can cover related languages.

Here we focus on the pair of Czech and Slovak, languages so similar that only morphological analysis and generation are needed. A successful system for this pair was already implemented: Česílko [10].

With the current surge of statistical MT systems and also large amounts of parallel data becoming available, we believe that the available systems should be re-evaluated to verify whether the old assumption still holds, i.e. whether tailor-made rule-based systems are still the best choice for closely related languages.

To the best of our knowledge, the only system ever evaluated on the Czech-to-Slovak pair was Česílko: TRADOS "match" between the output of the system and its manually post-edited version reached 90% but it is hard to interpret this single figure. In this paper, we try to fill the gap and provide a multifaceted evaluation of several MT systems for our language pair.

After a brief description of the available systems and parallel data (Section 2), we discuss the known problems of MT evaluation and also the problems specific for closely related languages (Section 3). Empirical results (Section 4) of the test include automatic as well as manual evaluation techniques and error analysis.

2 Available Resources

2.1 Data Used

The domain of the translated text is a critical factor in achievable translation quality and thus in the result of the evaluation. We would like to provide a balanced view, so we are using texts from three different domains. We also train a statistical system (see below) on the data of two of the domains.[1]

- **JRC-Acquis[16]**[2] is a multi-parallel corpus created from legislative texts used by the European Union. As such, the texts are very repetitive and the syntax is rather formal and complex. We used the third version of the corpus and extracted the Czech-Slovak parallel data, including the sentence alignment as officially released.
  The corpus Acquis was already divided into testing, development and traning sets by Philipp Koehn[3], so we stick to this division. (We use the file `ac-test` for testing, `ac-dev` for tuning and the rest for the training).

- **Books.**
  Thanks to the Slovak Academy of Science, we acquired 118 parallel Czech-Slovak books which will later become part of a larger corpus. A new version of the corpus based on the books is currently being developed. We processed and automatically aligned these books and selected 39 books translated from Czech to Slovak with the highest quality of the alignment. The alignment quality of the

* This work has been supported by the grants EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), P406/10/P259, and MSM 0021620838.

[1] To avoid technical problems, we ignored all sentences over 40 tokens.

[2] http://optima.jrc.it/Acquis/

[3] http://matrix.statmt.org/test_sets/list

rest was not sufficient for the machine translation usage. Then we randomly selected and extracted 4000 lines (i.e. parallel sentences) as the test data and another 4000 lines for tuning. The rest of the books served as training data.

- **WMT [5]**[4] is a workshop that annually runs shared translation task of mainly news texts. The test sets are different each year and we used the test set from 2010. The workshop is nowadays one of the "standard" venues to test machine translation at. Unfortunately, there is no Slovak version of the data, so we decided to use this dataset only as a small sample, translating the first 50 sentences manually from the Czech version.

The sizes of our training and tuning sets are listed in Table 1.

Corpora	Training Sentences	Tuning Sentences
Acquis	708406	3148
Books	137027	3802

Table 1. Number of sentences used for training and testing, after the removal of sentences over 40 tokens.

2.2 Systems Tested

We are aware only of the following existing MT systems for Czech-Slovak:

- **Google Translate**[5] is an online statistical translation system. It is trained on unspecified amount and type of text. Because of this, it could be hard to predict the behavior of this system on a particular type of test set, it is even possible that there is an overlap between our test sets and Google training data. In general, Google is known to perform well on varied domains.
  Within our experiments, we found some problems with translation of larger texts in Google Translate. Only the beginnings of texts were usually translated when the input was too large. Therefore, we had to divide larger documents into smaller parts and translate them separately.

- **Česílko 1.0** [10] is the above mentioned system aimed at the translation between closely related languages. Among other language pairs, it supports Czech-Slovak language pair. It uses direct word-to-word translation.
  As Česílko 1.0 needs input in the 8-bit encoding ISO-8859-2, texts had to be converted from UTF8. The ISO-8859-2 is somewhat limited in the set of character supported so we approximated e.g. the curly quotation marks („“) with the plain ASCII ones (") This could have minor influence on the scores, both automatic as well as manual.

- **Česílko 2.0** [9] is a reimplementation of the original Česílko 1.0 with one significant difference. The source-side morphological analyzer does not include statistical tagger and therefore it produces ambiguous output. This leads to a completely different architecture of the system, as the transfer unit must be able to work with ambiguous input. The transfer unit itself does not disambiguate. A new unit is introduced for this task: a statistical ranker, which chooses the best one of transferred sentences according to a target language model; this is where the missing disambiguation of the tagger is compensated. A technical issue of the current implementation is that the ranking is performed on the whole sentences, so each additional ambiguity multiplies the number of outputs.
  We encountered several problems when running Česílko 2.0 on the test data. A minor issue was caused by an internal incompatibility: the built-in morphological analyzer unit was producing tags unknown to the transfer unit. These included phenomena less frequent in written texts like pronouns in second person, names with ambiguous gender (like "animate or inanimate masculine" or "not feminine"), adjectives in nominal form, and others. We opted for a simple work-around to obtain somewhat distorted output: we forced the morphological analyzer to avoid such output tags altogether. Most frequently, this ultimately leads to the word being unrecognized and passed to the output without any change. In 424 cases of the 3125 sentences in the Acquis test set, Česílko 2.0 produced no output at all. We substituted the original Czech sentence as a fallback.

To complement the set of existing systems, we provide two systems of our own, based on the same engine but trained on different training data:

- **Moses** [11] is an open-source statistical phrase-based translation system. While Moses supports additional source or target "factors" to explicitly handle additional linguistic annotation such as morphology, we used it only in the baseline configuration.
  We trained and tuned Moses on the Acquis training data and independently on the Books training

[4] http://matrix.statmt.org/test_sets/list

[5] http://translate.google.com

and tuning data, obtaining two different MT systems.

3 Evaluating MT Quality for Closely Related Languages

The evaluation of MT outputs is a rather difficult task and the research community is still actively discussing and testing various manual and many automatic methods, see e.g. [5]. For example, different types of *manual* scoring lead to different results: one system can be rather poor when the annotators are asked to rank candidate sentences from best to worst but very good when the annotators are short given machine-translated texts and are asked to answer a set of yes/no questions about the content [2].

We are not aware of any evaluation technique tailored to closely related languages. However, there are specifics of the task that seriously affect the reliability of the evaluation.

It is well acknowledged in translation studies that the text is different when directly written in a language and when it was translated to the language. The source language used for the translation also plays an important role. The current state of the art in machine translation is not advanced enough to be heavily affected by such differences (provided that all the systems are run under the same conditions), although the differences are indeed measurable [13, 14]. It is common to simply ignore what was the original source and target language.[6]

For closely related languages, the impact of the source language is much more pronounced. If a system is based on the assumption that the translation can be more or less word for word, it will be heavily penalized if the reference translation significantly deviates from the source. Such a deviation is very likely if both source and reference come from a third language. In our case, this concerns primarily the Acquis corpus where most of the texts were translated from an English original to Czech and Slovak independently.

The books test set matches our translation direction: all were Czech books translated to Slovak. The same holds for our 50 WMT sentences, but here the translation was not professional (and thus actually likely to be rather verbatim).

4 Empirical Evaluation

The quality of machine translation output can be done both automatically and manually. The automatic methods rely on one or more reference translations and somehow calculate the similarity to the reference. Traditionally, they are called “metrics”, despite not satisfying the formal properties of metrics. The main advantage is the speed and the deterministic nature that allows to check progress on a fixed test set. On the other hand, the particular implementation of the similarity can heavily affect the bias towards certain MT system types. Moreover, experiments show that the correlation between used evaluation metrics and human judgment may be weak in some cases (e.g. [6]) and the problem is even worse when automatic metrics are applied to languages with rich morphology [12] and languages with higher degree of word order freedom. In general, it is usually preferred to consider several independent MT metrics. Our automatic evaluation is given in Section 4.1.

Manual evaluation methods are labor-intensive, subjective (different judges score systems differently for various reasons, incl. different expertise in the source language) and not reproducible (the same judge can not reliably evaluate the same set of sentences several times). On the other hand, they are the only real benchmark. Again, there are many possible manual evaluation techniques out of which we use two: we rank the hypotheses of various systems indicating which of them is overall better (Section 4.2) and we also mark and count errors in MT outputs (Section 5).

4.1 Automatic Evaluation Metrics

We evaluated the five systems using the following automatic metrics: BLEU [15], NIST [7], METEOR [1], TER [17], TERp-A and TERP-TER. All these metrics are based on comparison of acquired translation (the hypothesis) and the reference translation. BLEU score in essence counts the number of n-grams that occur in both translated sentence and the reference translation (and also takes care of the overall length of the output). NIST is very similar to BLEU score but the information gain of each n-gram is considered, auxiliary words thus tend to become less important. METEOR is based on unigram precision and unigram recall metrics, emphasizing the recall. TER (Translation Error Rate) gives number of edit operations that are needed to convert the hypothesis to the reference (word insertion, deletion, substitution and the movement of a sequence of words to a different position in the sentence). TERp-A and TERP-TER are based on TER score and they extend it by allowing e.g. paraphrases.

[6] Interestingly, the WMT evaluation campaigns [5], perhaps somewhat unintended, average out the phenomenon: each examined language contributes only a portion of the whole test set, the portion is then translated to all other examined languages.

Acquis	BLEU	NIST	METEOR	TER	TERpa	TERpter
Česílko 1.0	0.33	7.39	0.54	0.55	0.61	0.47
Česílko 2.0	0.15	4.71	0.35	0.75	0.88	0.64
Google Translate	**0.57**	**10.28**	**0.74**	**0.36**	**0.38**	**0.30**
Moses-Acquis	0.46	8.82	0.66	0.48	0.47	0.38
Moses-Books	0.22	5.89	0.49	0.66	0.72	0.56
WMT10 (50 sents)	BLEU	NIST	METEOR	TER	TERpa	TERpter
Česílko 1.0	0.68	7.89	0.85	0.16	0.22	0.14
Česílko 2.0	0.29	5.24	0.57	0.44	0.59	0.38
Google Translate	**0.78**	**8.43**	**0.90**	**0.11**	**0.14**	**0.09**
Moses-Acquis	0.38	5.88	0.63	0.39	0.50	0.35
Moses-Books	0.57	7.12	0.78	0.25	0.32	0.22
Books	BLEU	NIST	METEOR	TER	TERpa	TERpter
Česílko 1.0	0.39	8.73	0.65	0.44	0.50	0.38
Česílko 2.0	0.20	6.07	0.46	0.65	0.72	0.53
Google Translate	0.45	9.44	0.70	**0.41**	0.44	0.35
Moses-Acquis	0.18	5.63	0.42	0.64	0.75	0.55
Moses-Books	**0.47**	**9.74**	**0.71**	0.42	**0.41**	**0.33**

Table 2. Results of automatic scores on three test sets. Best results in bold.

All metrics were applied case insensitive. The metrics TER, TERp and METEOR are language dependent, but they do not support Slovak yet. As a substitute, we use Czech for METEOR and English for TER and TERp. These metrics could also benefit from WordNet installation to increase the coverage of words matched between the hypothesis and the reference. Unfortunately, there is no Slovak WordNet and the Czech WordNet is not easy to obtain. We used the English WordNet to satisfy the technical requirement, but we acknowledge that it can not possibly help in the evaluation. All such metrics are therefore at their baseline performance levels.

We used the default setting for all the metrics except METEOR where we explicitly asked for text normalization.

All systems were evaluated using three testing sets: Acquis (3125 sentences), Books (3860 sentences) and WMT10 (50 sentences).

Table 2 documents that all the automatic metrics provide the same picture. METEOR and the various versions of TER were shown to correlate better with humans than e.g. BLEU but the lack of Slovak resources inhibits their advantage.

As expected, the match between the system and the test set is critical. Google Translate performs very well on all the datasets.

Moses seems to work very well on the same type of data as it was trained on: both for Acquis and Books, the corresponding instance of Moses is the first or the second system. However, we should note that Moses trained on Books had a slight advantage over other systems: when selecting the test sets from the Books corpora, we extracted random *sentences* but we left other parts of the same document in the training data. Therefore, Moses had access to the exact terms and names used in each of the documents. It could have even happened that the same sentence appeared several times in the full corpus and some of the copies became part of the test set while others remained in the training data.

A more serious issue is the one discussed above in Section 3. Česílko had a big disadvantage on the Acquis test set, because both the source Czech and the target Slovak come from English. The training data for Moses-Acquis were created using the same procedure, so they are likely to account for the divergence caused by the third language.

The small WMT10 set is different: it is out of the domain for both variants of Moses and moreover it was translated word for word, in line with the algorithm of Česílko.

4.2 Manual Ranking of Systems

We carried out manual ranking of MT outputs using the same procedure as in WMT evaluation campaigns: Each annotator is given many “screens” or “hits” with the source sentence and the hypotheses of up to five systems. We did not provide any reference translation, because Czech is understood by our Slovak annotators. The task at each hit is to rank the hypotheses by assigning numbers to each hypothesis. Ties are allowed. We collected 3 independent judgments.

The official WMT interpretation is based on *pairs* of judgments. The 5 systems ranked at once imply 10 pairwise comparisons per hit. For each system, we divide the number of (strict) wins by the total number

Acquis	> others	>= others	> all in hit	>= all in hit
Česílko 1.0	49.0%	**87.2%**	**19.8%**	**71.6%**
Česílko 2.0	4.6%	29.6%	0.0%	21.2%
Google Translate	**51.5%**	86.5%	19.0%	70.9%
Moses-Acquis	41.3%	81.3%	11.9%	63.1%
Moses-Books	24.6%	41.7%	2.6%	17.1%
WMT10 (50 sents)	> others	>= others	> all in hit	>= all in hit
Česílko 1.0	52.0%	76.4%	19.5%	56.1%
Česílko 2.0	16.7%	44.7%	0.0%	11.4%
Google Translate	**65.8%**	**86.3%**	**43.6%**	**76.9%**
Moses-Acquis	13.3%	36.2%	0.0%	5.7%
Moses-Books	37.4%	69.1%	12.2%	41.5%
Books	> others	>= others	> all in hit	>= all in hit
Česílko 1.0	50.6%	79.1%	15.6%	53.7%
Česílko 2.0	16.3%	38.1%	1.2%	11.8%
Google Translate	**61.0%**	**85.7%**	**33.8%**	**68.8%**
Moses-Acquis	13.2%	34.1%	0.7%	9.4%
Moses-Books	55.3%	78.7%	28.9%	59.6%

Table 3. Manual evaluation on our three test sets.

System	Rank
A	2
B	1
C	2

Implied Pairwise Comparisons
A<B
A=C
B>C

Fig. 1. An example of manual ranking of three systems in one "hit". The smaller the rank, the better the system. The system B gains two of two possible points in the official score ">= others" because it won both its two pairwise comparisons. It gets one of one possible point in ">= all in hit". The systems A and C each gain one of two possible points in ">= others" and no point in ">= all in hit".

of pairwise comparisons it took part in to obtain the official percentages called "> others" and ">= others". We add one more interpretation: we measure the number of *hits* where the system was the only winner (>) or one of several winners (>= all in hit). An example is given in Figure 1.

Based on the manual ranking (Table 3), Google performs indeed very well but Moses often loses compared to Česílko. It is only the Books test set where Moses scores slightly better but its performance may be overestimated due to the test set selection as described above. On the Acquis set, Česílko 1.0 even won under some of the interpretations, e.g. being among the winners in a hit.

From this we conclude that, unless Google-sized text data (both parallel and especially monolingual) are available, for closely-related languages (rule-based) systems implementing a shallow transfer are much more robust to domain effects than statistical MT systems. Often, this makes their output simply better than the output of SMT.

However, as indicated by the very bad results for Česílko 2.0, regular testing of such systems is critical. The new version of Česílko performs far worse than the ten year old one, at least in terms of software stability. It is hard to draw a conclusion on translation quality given that 14% of sentences were not translated at all.

5 Error Analysis

To complement the ranking and refine the error analysis, we manually flagged errors in the fifty sentences of our WMT test set following the error classification by [18] and further examined by [4].

As we see in Table 4, the number of errors corresponds to the overall ranking of the systems on the WMT set.

Moses trained on the Acquis data had a lot of problems with vocabulary on the WMT corpora. As expected, training Moses on Books, a more diverse dataset, improved the lexical choice and the number of translated words. The number of errors in word form choice is similar for Česílko 1.0 and both instances of Moses. Google is able to outperform this using a large language model. Moses factored setups (and larger Slovak monolingual data) could be also used to improve the form choice [3].

Most of the errors produced by Česílko 1.0 were caused by wrong form selection. Altogether, the word-for-word translation works very well in most of cases. Sometimes, slightly different word order would be preferred, but the approximation taken by Česílko is acceptable. The coverage of the lexicon seems to be good

Type of	Česílko	Česílko	Google	Moses	Moses
Error	1.0	2.0	Translate	Acquis	Books
Bad Disambiguation	3	11	6	17	12
Bad Lex. Choice	2	6	1	22	1
Bad Negation	0	1	0	0	0
Total Bad Word Sense	5	18	7	39	13
Missing Aux. Word	0	0	1	2	0
Missing Content Word	0	0	0	0	1
Total Missed Words	0	0	1	2	1
Bad Word Form	40	83	30	39	42
Extra Word	0	0	0	7	4
Untranslated Word	26	156	3	105	41
Total Serious Errors	71	257	41	192	101
Bad Word Order (Close)	0	0	0	8	0
Bad Word Order (Distant)	0	0	0	2	0
Bad Punctuation	0	0	1	1	0
Bad Letter Case	0	0	0	38	13
Total Errors	71	257	**42**	241	115

Table 4. Counts of various types of errors on the WMT test set (50 sent.)

enough, occasional gaps still result in an untranslated word and surprisingly, we have also seen badly generated (i.e. non-existing) Slovak words.

Česílko 2.0 proved to be in a very bad shape and not a solid representative of the shallow transfer.

6 Conclusion and Future Research

We prepared three test sets for the evaluation of Czech-Slovak translation and evaluated five MT systems both using automatic and manual evaluation techniques. The underlying question was whether shallow-transfer (rule-based) systems are still appropriate for closely related languages, given the large amounts of parallel texts available.

Aside from discussing the issues of evaluating such shallow-transfer systems, we confirmed the hypothesis: Česílko performed better than our statistical systems based on Moses. However, with huge amounts of data available, we expect statistical systems to dominate MT of closely related languages as well. Google Translate won in nearly all our evaluations.

For Czech-Slovak translation, we would like to improve Česílko 1.0 in areas identified by our error analysis: the lexicon coverage and esp. the morphological errors in the output.

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (2005)
2. Berka, J., Černý, M., Bojar, O.: Quiz-Based Evaluation of Machine Translation. Prague Bulletin of Mathematical Linguistics 95 (Mar 2011)
3. Bojar, O.: English-to-Czech Factored Machine Translation. In: Proceedings of the Second Workshop on Statistical Machine Translation. pp. 232–239. Association for Computational Linguistics, Prague, Czech Republic (June 2007)
4. Bojar, O.: Analyzing Error Types in English-Czech Machine Translation. Prague Bulletin of Mathematical Linguistics 95 (Mar 2011)
5. Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., Zaidan, O.: Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In: Proc. of WMT10 and MetricsMATR. pp. 17–53. ACL, Uppsala, Sweden (July 2010), revised August 2010
6. Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the Role of BLEU in Machine Translation Research. In: Proceedings of the EACL'06 (2006)
7. Doddington, G.: Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In: Proceedings of the second international conference on Human Language Technology Research (2002)
8. Hajič, J.: RUSLAN: an MT system between closely related languages. In: Proceedings of the third conference on European chapter of the Association for Computational Linguistics. pp. 113–117. Association for Computational Linguistics (1987)
9. Homola, P., Kuboň, V., Vičič, J.: Shallow transfer between slavic languages. In: Proceedings of Balto-Slavonic Natural Language Processing. pp. 219–232. Polska Akademia Nauk, Kraków, Poland (2009)
10. Hric, J., Hajič, J., Kuboň, V.: Machine Translation of Very Close Languages. Proc. of the 6th Applied Natural Language Processing Conference pp. 7–12 (2000)
11. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W.,

Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. pp. 177–180. ACL, Prague (June 2007)

12. Kos, K., Bojar, O.: Evaluation of machine translation metrics for czech as the target language. Prague Bulletin of Mathematical Linguistics 92 (2009)
13. Kurokawa, D., Goutte, C., Isabelle, P.: Automatic Detection of Translated Text and its Impact on Machine Translation. In: Proc. of MT Summit XII (2009)
14. Lembersky, G., Ordan, N., Wintner, S.: Language Models for Machine Translation: Original vs. Translated Texts. Abstract at Machine Translation and Morphologically-Rich Languages. Research Workshop of the Israel Science Foundation University of Haifa, Israel (Jan 2011)
15. Papineni K., Roukos S., Ward T, Zhy W.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
16. Ralf, S., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006). pp. 2142–2147. ELRA (2006)
17. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas (2006)
18. Vilar, D., Xu, J., D'Haro, L.F., Ney, H.: Error Analysis of Machine Translation Output. In: International Conference on Language Resources and Evaluation. pp. 697–702. Genoa, Italy (May 2006)