Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-58

Improving SMT by Using Parallel Data of a Closely Related Language

Petra Galuščáková [1] and Ondřej Bojar
Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Czech Republic

Abstract. The amount of training data in statistical machine translation critically affects translation quality. In this paper, we demonstrate how to increase translation quality for one language pair by introducing parallel data from a closely related language. Specifically, we improve English→Slovak translation using a large Czech–English parallel corpus and a shallow MT system for Czech→Slovak translation. Several options are explored to identify the best possible configuration.

We also present our two contributions to available data resources, namely the English–Slovak parallel corpus and the Slovak variant of the WMT 2011 test set.

Keywords. machine translation, parallel corpora, English-to-Slovak translation

Introduction

For closely related languages such as Czech–Slovak [1,2], Catalan–Spanish [3], Ukrainian–Russian [4] and possibly also Latvian–Lithuanian [5], remarkably good translation results can be achieved using relatively simple machine translation (MT) techniques.

If one of the related languages is under-resourced, the translation quality to or from that language into a third language may be improved by using data from the other related language. In our case we experimented with English→Slovak translation utilizing the MT system "Česílko" and a large Czech–English parallel corpus. We confirmed this expectation that pivoting through a related language is helpful and we also provided a description of a particular combination technique that worked best in our setting.

1. Related Work

The concept of using pivot or intermediate languages to improve MT quality has been widely studied. Babych et al. [4] show on a set of commercial MT systems that the pivoting is especially helpful if the pivot language is closely related to the source or target language (so this translation is not largely distorted) and when only a small amount of parallel data is available for the source or target language. We confirm this observation on another language pair and using a state-of-the-art phrase-based statistical MT system.

[1]Corresponding Author: Petra Galuščáková, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Malostranské náměstí 25, 11800 Prague, Czech Republic; E-mail: galuscakova@ufal.mff.cuni.cz.

A multitude of work on pivoting is available for statistical approaches to MT [6,7,8,9,10,11,12]. One method of using a pivot language is to combine the models (phrase tables) of two translation systems (from the source to pivot language and from pivot to target language). The pivot language or several pivot languages are usually used to *extend* the set of phrase pairs. Chen et al. [13] use a variant called "triangulation", where the MT systems based on different pivot languages have to *agree* on the translation. This has the positive effect of considerably reducing the phrase table size while increasing the translation quality.

Some authors prefer to work with MT outputs rather than the underlying models. They use either a method called *cascading* – combining the lists of the best translations – or they create artificial parallel data. The latter is usually achieved by automatically translating the pivot side of a parallel corpus and training the machine translation system on this half-synthetic parallel corpus.

Our experiments tend to be similar to some of the techniques used by Wu and Wang [11], who provide a comparison of several pivoting methods. The article discusses Chinese to Spanish translation using English as the pivot language. In contrast, we used a pivot language for the translation between closely related languages and so the improvement is expected to be more pronounced. Henríquez et al. [14] present a comparison of the approach based on pseudo corpus creation and concatenation of translation systems. Both of these methods are also used in our work.

Also existing are methods especially intended for translation between closely related languages which include e.g. transliteration [15,16], a method in which sequences of characters are processed rather than sequences of words.

The translation system Česílko, which we use in our experiments, was originally proposed for pivoting via Czech into close Slavic languages [2], but no experiments to support the authors' expectations have yet been performed according to our knowledge.

2. MT Systems Used

We use the following MT systems.

Česílko 1.0 is a stand-alone machine translation system intended for the translation between closely related languages. The version 1.0[2] we use supports only the Czech→Slovak pair. The system consists of a morphological analyzer, a statistical tagger, a simple dictionary (supporting also multi-word entries) for the transfer and morphological generation of Slovak. The system benefits from the closeness of the languages in question – e.g. it does not change word order during the translation. Česílko was chosen because it performed well in a comparison of several cs→sk translation systems [17]. While Česílko was not always best, it was fairly robust to various input text types.

Moses [18] is an open source statistical machine translation system. It was used as the baseline direct en→sk translation as well as for the various configurations of pivoting.

[2] http://hdl.handle.net/11858/00-097C-0000-0006-AAFE-A

Table 1. Sizes of our training corpora. The Slovak part of CzEng was created by Česílko as an automatic translation of the Czech portion.

	CzEng			En–Sk Corpus	
	English	Czech	Slovak (MT)	English	Slovak
Sentences	7.15 mil	7.15 mil	7.15 mil	2.46 mil	2.46 mil
Tokens	85.09 mil	72.86 mil	72.96 mil	52.09 mil	46.81 mil

3. Data Used

For English–Czech, training data were already available in the corpus CzEng[3]. We used version 0.9 and avoided the development and evaluation sections.

For English–Slovak and for the evaluation set, we prepared our own data collections, see below.

3.1. English-Slovak Parallel Corpus

The English–Slovak corpus was compiled from freely available sources: Acquis[4] version 3.0, European Commission Website[5] and parts of OPUS Corpus [19] (EMEA, EUconst, KDE4 and PHP). We did not use OPUS subtitles because we needed better quality data.

The corpus was further enlarged with data from the Journal of European Union and Europarl corpus. Then it was morphologically annotated by the Slovak Academy of Science and it is now freely available[6]. We did not use the morphological annotation in the experiments reported here.

3.2. Summary of Training Data

Table 1 summarizes the sizes of our training corpora. As outlined above, we automatically translated the Czech side of CzEng into Slovak using Česílko, which resulted in 72.96 million synthetic Slovak tokens, considerably more than the 46.81 million tokens of real Slovak in the En–Sk corpus.

3.3. Evaluation Data for {English, Czech, German, Spanish, French}–Slovak

Our test set is derived from the WMT 2011 shared task[7]. This test set consists of newspaper articles covering a broad range of topics. The test set is multi-parallel, available in Czech, English, German, Spanish and French. The source languages of the news articles differ – each article comes from one of the five languages and it is translated sentence by sentence to all the other languages.

We extended the dataset to include Slovak version[8]. The complete set was translated by eight translators, all of them were native Slovak speakers. Since Slovak speakers have a very good knowledge of Czech, the most reliable way of acquiring the translation was

[3] http://ufal.mff.cuni.cz/czeng/czeng09/
[4] http://langtech.jrc.it/JRC-Acquis.html
[5] http://ec.europa.eu/index_en.htm
[6] http://hdl.handle.net/11858/00-097C-0000-0006-AADF-0
[7] http://www.statmt.org/wmt11/
[8] http://hdl.handle.net/11858/00-097C-0000-0006-AADA-9

Table 2. Some translation inconsistencies in WMT11 Czech and English test set.

Czech	Gloss of Czech	English
Majitelé psů, Kristina Rickard a David Peek, nebyli zastiženi, aby **se mohli k případu vyjádřit**.	The dogs owners, Kristina Rickard and David Peek, could not be reached for **comment**.	The dogs owners, Kristina Rickard and David Peek, could not be reached for **comment Thursday**.
Nebyl od toho daleko.	**He was not far from it.**	**Far from it**.
Jak připomíná před členy parlamentu UMP ve **středu** večer v Elysejském paláci, perspektiva "dvou mandátů" zasela **málo** horečný stav.	The fact that on **Wednesday** night at the Elysée before UMP deputies he called up the prospect of "two mandates" sowed **little** feverish reflections.	The fact that on **Tuesday** night at the Elysée before UMP deputies he called up the prospect of "two mandates" sowed **some** feverish reflections.

Table 3. Sizes of our test sets.

	English	Czech	Slovak
Sentences	3 003	3 003	3 003
Tokens	77 086	68 108	63 730

to translate the Czech version into Slovak. We also decided to provide the English version of the text to the translators to help the translator to better understand the text, especially in ambiguous cases.

Surprisingly, many discrepancies between the English and Czech sentences in the original WMT data were found. In several cases, some entire sentences have different meanings in Czech versus English. Some of the problems are illustrated in Table 2.

Because of the multiple original languages of the texts, it is difficult to determine which version of the sentence is the correct one. Since we wanted to use the test set primarily for the en→sk MT evaluation, we encouraged translators to prefer the English version in problematic cases. Finally, the translations were automatically checked using a few scripts: e.g. multiple spaces and incorrect punctuation were corrected.

Table 3 summarizes the sizes of the relevant versions of the test set.

4. Setups Examined

We studied the following setups:

Direct Translation. Statistical translation system Moses is trained and tuned on English–Slovak parallel data. The resulting model is used for direct English→Slovak translation.

Moses+Česílko. Czech is used as a pivot language, simple *cascading* is applied. Moses is trained and tuned on the English–Czech corpus. The resulting model is used for English→Czech translation, the output of which is further translated into Slovak by Česílko.

Česílko+Moses. The synthetic corpus is created. The Czech part of the English–Czech corpus is automatically translated by Česílko into Slovak. Moses is trained and

tuned on this synthetic parallel corpus and the model is used for English→Slovak translation.

Česílko+Moses+Direct A combination of the previous first and third options. The training data are acquired as the concatenation of the manual English–Slovak corpus (as used in Direct Translation) and the synthetic English–Slovak corpus from Česílko+Moses. This combined corpus is used for training of Moses and the model is used for English→Slovak translation.

5. Experimental Results

We evaluate all our experiments using two automatic evaluation metrics that compare the MT output to the reference translation: BLEU [20] and TER [21]. Note that all experiments in this work were performed on lower case data.

We first mention two tricks and then provide the comparison of the four main setups.

5.1. Minor Tricks in Tuning Moses

Phrase-based MT uses a complex processing pipeline where many little deviations from the default can lead to improvements in translation quality. We used the following two tricks.

For the evaluation of the tricks, we split the WMT 2011 test into two halves. The first half served as our tuning set and the second half was used for evaluation.

5.1.1. Simple Stemming for Word Alignment

Only the first 4 letters of each word in both source and target languages were used for word alignment to overcome data sparseness. (The translation model is then obviously based on word forms, not these simple stems.) Table 4 compares this approach for English-to-Slovak to the default where fully inflected word forms are used throughout the processing pipeline.

5.1.2. Synthetic Tuning Data

The last step of Moses training pipeline is the tuning of model weights on an independent set of sentences. We examined whether the reference translation for English→Slovak should be rather manually translated or whether the automatic Slovak obtained by Česílko from Czech would yield better results.

Again, we use the split WMT 2011 test set where the first half serves for tuning, either in its manual Slovak version, or an automatic version obtained by Česílko from the Czech side. The second half (always manual translation) was used for evaluation.

Table 4. English→Slovak translation when whole word forms were used and when only the first four characters were used to obtain word alignments. Empirical 95% confidence intervals are in brackets.

Preprocessing for word alignment	BLEU	TER
Word Form	0.1165 [0.1104,0.1227]	0.7143 [0.7052,0.7143]
First 4 Characters	0.1211 [0.1151,0.1275]	0.7071 [0.6981,0.716]

Table 5. Comparison of scores achieved by the same MT system when tuned on either a manually or automatically translated tuning set. Empirical 95% confidence intervals are in brackets.

Reference of the tuning set	BLEU	TER
Automatic	0.1273 [0.1215, 0.1332]	0.6880 [0.6794, 0.6966]
Manual	0.1261 [0.1201, 0.1319]	0.6888 [0.6803, 0.6977]

Table 6. Scores of English→Slovak translation achieved using several methods with the support of English–Czech data. Empirical confidence intervals are in brackets.

	BLEU	TER
Direct Translation	0.1083 [0.1039, 0.1125]	0.7248 [0.7189, 0.7314]
Moses+Česílko	0.1131 [0.1089, 0.1171]	0.7111 [0.7049, 0.7171]
Česílko+Moses	0.1189 [0.1143, 0.1230]	0.7049 [0.6986, 0.7113]
Česílko+Moses+Direct	0.1261 [0.1213, 0.1305]	0.6914 [0.6851, 0.6979]

We used the English–Czech corpus translated to Slovak for training. The final scores achieved using the automatically translated tuning data were slightly better than the results of the experiment which used manually translated data, see Table 5.

This result may be caused by the properties of Česílko and BLEU evaluation metric. Česílko translates word by word and does not change the word order. This could lead to the higher scores calculated by BLEU. In any case, the difference is not significant.

Based on this initial experiment, we opted for a larger tuning set (which has been reported to help) and automatically translated the WMT 2010 test set from Czech into Slovak using Česílko. This synthetic tuning set was used for all the main experiments while the whole WMT 2011 test set (with manual Slovak) was reserved for testing only.

5.2. Pivoting Experiments

The results of our main experiments are tabulated in Table 6. Direct Translation is significantly worse than the results of all the other translation schemes. This means that we are able to achieve better en→sk results by using any of our suggested techniques employing English–Czech data. Because the English–Slovak corpus is almost three times smaller than the English–Czech corpus, this result is not surprising.

The result of Česílko+Moses, in which the English–Czech corpus is translated into Slovak and then used for training, performs significantly better than the converse Moses+Česílko when Moses operates on English→Czech and the resulting Czech is then translated into Slovak by Česílko. Our approach based on the creation of a synthetic parallel corpus thus outperformed the simple cascading method of two systems in sequence. The best result was achieved in the fourth case – when both corpora, the smaller manual English–Slovak and the larger English–Czech automatically translated to Slovak, were used.

Because the testing set is also available in Czech, we also tried to translate the Czech part of the corpus into Slovak using Česílko. The BLEU score for the cs→sk translation of the same testing set is 42.45, with the confidence interval [41.67,43.18]. This high score is not surprising. Česílko preserves the word order and the translators may have

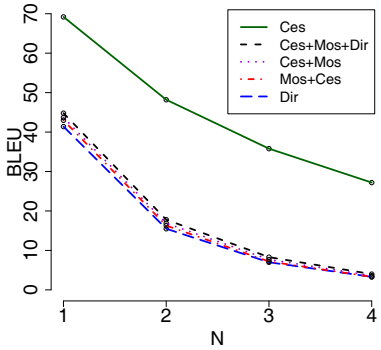

Figure 1. BLEU score components (various n-gram precisions).

pursued the same approach because they were also translating from Czech. BLEU gives then a high credit to the matching n-grams.

Similarly to Babych et al. [4], we examined the components of BLEU scores for various n-grams, see Figure 1. The tendency is the same for all en→sk translations, the n-gram precision decreases exponentially with n. For a translation between closely related languages, from Czech into Slovak, the decrease is not that pronounced. This result agrees with the observations of Babych et al.: a linear decrease of the n-gram precision for closely related languages and an exponentional decrease in the case of distant languages.

6. Conclusion

We have examined several techniques for improving the quality of English→Slovak machine translation by employing language resources of a closely related language, namely Czech.

We confirmed our expectation that pivoting via a closely related language performs well. In our experiments, the method based on the creation of a synthetic parallel corpus by translating the Czech side of an English–Czech parallel corpus gave superior results comparing to the simple cascading of the en→cs and cs→sk translation systems. The best result was obtained using all available data: the parallel corpus for the direct en→sk translation as well as the synthetic corpus constructed using Czech→Slovak shallow MT.

Acknowledgments

This work was supported by the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003+7E11051 of the Czech Republic), the Czech Science Foundation grant P406/10/P259 and SVV project number 265 314.

References

[1] J. Hajič, V. Kuboň, and J. Hric, Machine Translation of Very Close Languages, In *Proc. of the 6th Applied Natural Language Processing Conference*, Seattle, Washington (2000), 7–12.

[2] J. Hajič, P. Homola, and V. Kuboň, A Simple Multilingual Machine Translation System, In *Proc. of MT Summit IX*, New Orleans, USA (2003), 157–164.

[3] J. A. Alonso, Machine Translation for Catalan-Spanish: The Real Case for Productive MT, In *Proc. of the tenth Conference on European Association of Machine Translation*, Budapest, Hungary (2005), 23–26.

[4] B. Babych, A. Hartley, and S. Sharoff, Translating from Under-resourced Languages: Comparing Direct Transfer against Pivot Translation, In *Proc. of MT Summit XI*, Copenhagen, Denmark (2007), 29–35.

[5] M. Khalilov, L. Pretkalnina, N. Kuvaldina, and V. Pereseina, SMT of Latvian, Lithuanian and Estonian Languages: A Comparative Study, In *Proc. of the 4th Internation Conference on Human Language Technologies - the Baltic Perspective (HLT'10)*, Riga, Latvia (2010), 117–124.

[6] T. Gollins and M. Sanderson, Improving Cross Language Retrieval with Triangulated Translation, In *Proc. of ACM SIGIR*, New Orleans, USA (2001), 90–95.

[7] T. Cohn and M. Lapata, Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora, In *Proc. of ACL*, Prague, Czech Republic (2007), 728–735.

[8] M. Utiyama and H. Isahara, A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation, In *Proc. of HLT-NAACL*, Rochester, New York, USA (2007), 484–491.

[9] N. Bertoldi, M. Barbaiani, M. Federico, and R. Cattoni, Phrase-Based Statistical Machine Translation with Pivot Languages, In *Proc. of IWSLT Evaluation Campaign on Spoken Language Translation*, Honolulu, Hawaii, USA (2008), 143–149.

[10] M. Paul, H. Yamamoto, E. Sumita, and S. Nakamura, On The Importance of Pivot Language Selection for Statistical Machine Translation, In *Proc. of HLT-NAACL (Short Papers)*, Boulder, Colorado, USA (2009), 221–224.

[11] H. Wu and H. Wang, Revisiting pivot language approach for machine translation, In *Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, Suntec, Singapore (2009), 154–162.

[12] J. Tiedemann, Character-Based Pivot Translation for Under-Resourced Languages and Domains, In *Proc. of EACL*, Avignon, France (2012), 141–151.

[13] Y. Chen, M. Kay, and A. Eisele, Intersecting Multilingual Data for Faster and Better Statistical Translations, In *Proc. of HLT/NAACL*, Boulder, Colorado (2009), 128–136.

[14] C. Henríquez, M. R. Costa-jussà, R. E. Banchs, L. Formiga, and J. B. Mariño, Pivot Strategies as an Alternative for Statistical Machine Translation Tasks Involving Iberian Languages, In *Workshop on ICL NLP Tasks*, Huelva, Spain (2011), 22–27.

[15] D. Vilar, J.-T. Peter, and H. Ney, Can We Translate Letters?, In *Proc. of ACL WMT07*, Prague, Czech Republic (2007), 33–39.

[16] P. Nakov and H. T. Ng, Improved Statistical Machine translation for resource-poor languages using related resource-rich Languages, In *Proc. of EMNLP*, Singapore (2009), 1358–1367.

[17] O. Bojar, P. Galuščáková, and M. Týnovský, Evaluating Quality of Machine Translation from Czech to Slovak, In *Information Technologies – Applications and Theory*, Košice, Slovakia (2011), 3–9.

[18] 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, In *Proc. of ACL Demo and Poster Sessions*, Prague, Czech Republic (2007), 177–180.

[19] J. Tiedemann, Parallel Data, Tools and Interfaces in OPUS, In *Proc. of LREC*, Istanbul, Turkey (2012),.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, In *Proc. of ACL 2002*, Philadelphia, Pennsylvania (2002), 311–318.

[21] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, A Study of Translation Edit Rate with Targeted Human Annotation, In *Proc. of AMTA*, Cambridge, Massachusetts, USA (2006), 223–231.