CUni Multilingual Matrix in the WMT 2013 Shared Task

Karel Bílek **Daniel Zeman**
Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, CZ-11800 Praha, Czechia
kb@karelbilek.com, zeman@ufal.mff.cuni.cz

Abstract

We describe our experiments with phrase-based machine translation for the WMT 2013 Shared Task. We trained one system for 18 translation directions between English or Czech on one side and English, Czech, German, Spanish, French or Russian on the other side. We describe a set of results with different training data sizes and subsets. For the pairs containing Russian, we describe a set of independent experiments with slightly different translation models.

1 Introduction

With so many official languages, Europe is a paradise for machine translation research. One of the largest bodies of electronically available parallel texts is being nowadays generated by the European Union and its institutions. At the same time, the EU also provides motivation and boosts potential market for machine translation outcomes.

Most of the major European languages belong to one of three branches of the Indo-European language family: Germanic, Romance or Slavic. Such relatedness is responsible for many structural similarities in European languages, although significant differences still exist. Within the language portfolio selected for the WMT shared task, English, French and Spanish seem to be closer to each other than to the rest.

German, despite being genetically related to English, differs in many properties. Its word order rules, shifting verbs from one end of the sentence to the other, easily create long-distance dependencies. Long German compound words are notorious for increasing out-of-vocabulary rate, which has led many researchers to devising unsupervised compound-splitting techniques. Also, uppercase/lowercase distinction is more important because all German nouns start with an uppercase letter by the rule.

Czech is a language with rich morphology (both inflectional and derivational) and relatively free word order. In fact, the predicate-argument structure, often encoded by fixed word order in English, is usually captured by inflection (especially the system of 7 grammatical cases) in Czech. While the free word order of Czech is a problem when translating to English (the text should be parsed first in order to determine the syntactic functions and the English word order), generating correct inflectional affixes is indeed a challenge for English-to-Czech systems. Furthermore, the multitude of possible Czech word forms (at least order of magnitude higher than in English) makes the data sparseness problem really severe, hindering both directions.

Most of the above characteristics of Czech also apply to Russian, another Slavic language. Similar issues have to be expected when translating between Russian and English. Still, there are also interesting divergences between Russian and Czech, especially on the syntactic level. Russian sentences typically omit copula in the present tense and there is also no direct equivalent of the verb “to have”. Periphrastic constructions such as “there is XXX by him” are used instead. These differences make the Czech-Russian translation interest-

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 85–91,
Sofia, Bulgaria, August 8-9, 2013

ing as well. Interestingly enough, results of machine translation between Czech and Russian has so far been worse than between English and any of the two languages, language relatedness notwithstanding.

Our goal is to run one system under as similar conditions as possible to all eighteen translation directions, to compare their translation accuracies and see why some directions are easier than others. The current version of the system does not include really language-specific techniques: we neither split German compounds, nor do we address the peculiarities of Czech and Russian mentioned above.

In an independent set of experiments, we tried to deal with the data sparseness of Russian language with the addition of a backoff model with a simple stemming and some additional data; those experiments were done for Russian and Czech|English combinations.

2 The Translation System

Both sets of experiments use the same basic framework. The translation system is built around Moses[1] (Koehn et al., 2007). Two-way word alignment was computed using GIZA++[2] (Och and Ney, 2003), and alignment symmetrization using the *grow-diag-final-and* heuristic (Koehn et al., 2003). Weights of the system were optimized using MERT (Och, 2003). No lexical reordering model was trained.

For language modeling we use the SRILM toolkit[3] (Stolcke, 2002) with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998).

3 General experiments

In the first set of experiments we wanted to use the same setting for all language pairs.

3.1 Data and Pre-processing Pipeline

We applied our system to all the ten official language pairs. In addition, we also experimented with translation between Czech on one side and German, Spanish, French or Russian on the other side. Training data for these additional language pairs were obtained by combining parallel corpora of the officially supported pairs. For instance, to create the Czech-German parallel corpus, we identified the intersection of the English sides of Czech-English and English-German corpora, respectively; then we combined the corresponding Czech and German sentences.

We took part in the constrained task. Unless explicitly stated otherwise, the translation model in our experiments was trained on the combined News-Commentary v8 and Europarl v7 corpora.[4] Note that there is only News Commentary and no Europarl for Russian. We were also able to evaluate several combinations with large parallel corpora: the UN corpus (English, French and Spanish), the Giga French-English corpus and CzEng (Czech-English). We did not use any large corpus for Russian-English. Table 1 shows the sizes of the training data.

Corpus	SentPairs	Tkns lng1	Tkns lng2
cs-en	786,929	18,196,080	21,184,881
de-en	2,098,430	55,791,641	58,403,756
es-en	2,140,175	62,444,507	59,811,355
fr-en	2,164,891	70,363,304	60,583,967
ru-en	150,217	3,889,215	4,100,148
de-cs	657,539	18,160,857	17,788,600
es-cs	697,898	19,577,329	18,926,839
fr-cs	693,093	19,717,885	18,849,244
ru-cs	103,931	2,642,772	2,319,611
Czeng			
cs-en	14,833,358	204,837,216	235,177,231
UN			
es-en	11,196,913	368,154,702	328,840,003
fr-en	12,886,831	449,279,647	372,627,886
Giga			
fr-en	22,520,400	854,353,231	694,394,577

Table 1: Number of sentence pairs and tokens for every language pair in the parallel training corpus. Languages are identified by their ISO 639 codes: cs = Czech, de = German, en = English, es = Spanish, fr = French, ru = Russian. Every line corresponds to the respective version of EuroParl + News Commentary; the second part presents the extra corpora.

The News Test 2010 (2489 sentences in each language) and 2012 (3003 sentences) data sets[5] were used as development data for MERT. BLEU scores reported in this paper were computed on the News Test 2013 set

[1] http://www.statmt.org/moses/
[2] http://code.google.com/p/giza-pp/
[3] http://www-speech.sri.com/projects/srilm/
[4] http://www.statmt.org/wmt13/translation-task.html\#download
[5] http://www.statmt.org/wmt13/translation-task.html

(3000 sentences each language). We do not use the News Tests 2008, 2009 and 2011.

All parallel and monolingual corpora underwent the same preprocessing. They were tokenized and some characters normalized or cleaned. A set of language-dependent heuristics was applied in an attempt to restore the opening/closing quotation marks (i.e. `"quoted"` → “quoted”) (Zeman, 2012).

The data are then tagged and lemmatized. We used the Featurama tagger for Czech and English lemmatization and TreeTagger for German, Spanish, French and Russian lemmatization. All these tools are embedded in the Treex analysis framework (Žabokrtský et al., 2008).

The lemmas are used later to compute word alignment. Besides, they are needed to apply “supervised truecasing” to the data: we cast the case of the lemma to the form, relying on our morphological analyzers and taggers to identify proper names, all other words are lowercased. Note that guessing of the true case is only needed for the sentence-initial token. Other words can typically be left in their original form, unless they are uppercased as a form of HIGHLIGHTING.

3.2 Experiments

BLEU scores were computed by our system, comparing truecased tokenized hypothesis with truecased tokenized reference translation. Such scores must differ from the official evaluation—see Section 3.2.4 for discussion of the final results.

The confidence interval for most of the scores lies between ±0.5 and ±0.6 BLEU % points.

3.2.1 Baseline Experiments

The set of baseline experiments were trained on the supervised truecased combination of News Commentary and Europarl. As we had lemmatizers for the languages, word alignment was computed on lemmas. (But our previous experiments showed that there was little difference between using lemmas and lowercased 4-character “stems”.) A hexagram language model was trained on the monolingual version of the News Commentary + Europarl corpus (typically a slightly larger superset of the target side of the parallel corpus).

3.2.2 Larger Monolingual Data

Besides the monolingual halves of the parallel corpora, additional monolingual data were provided / permitted. Our experiments in previous years clearly showed that the Crawled News corpus (2007–2012), in-domain and large, contributed significantly to better BLEU scores. This year we included it in our baseline experiments for all language pairs: translation model on News Commentary + Europarl, language model on monolingual part of the two, plus Crawled News.

In addition there are the Gigaword corpora published by the Linguistic Data Consortium, available only for English (5^{th} edition), Spanish (3^{rd}) and French (3^{rd}). Table 2 gives the sizes and Table 3 compares BLEU scores with Gigaword against the baseline. Gigaword mainly contains texts from news agencies and as such it should be also in-domain. Nevertheless, the crawled news are already so large that the improvement contributed by Gigaword is rarely significant.

Corpus	Segments	Tokens
newsc+euro.cs	830,904	18,862,626
newsc+euro.de	2,380,813	59,350,113
newsc+euro.en	2,466,167	67,033,745
newsc+euro.es	2,330,369	66,928,157
newsc+euro.fr	2,384,293	74,962,162
newsc.ru	183,083	4,340,275
news.all.cs	27,540,827	460,356,173
news.all.de	54,619,789	1,020,852,354
news.all.en	68,341,615	1,673,187,787
news.all.es	13,384,314	388,614,890
news.all.fr	21,195,476	557,431,929
news.all.ru	19,912,911	361,026,791
gigaword.en	117,905,755	4,418,360,239
gigaword.es	31,304,148	1,064,660,498
gigaword.fr	21,674,453	963,571,174

Table 2: Number of segments (paragraphs in Gigaword, sentences elsewhere) and tokens of additional monolingual training corpora. “newsc+euro” are the monolingual versions of the News Commentary and Europarl parallel corpora. “news.all” denotes all years of the Crawled News corpus for the given language.

Direction	Baseline	Gigaword
en-cs	0.1632	
en-de	0.1833	
en-es	0.2808	0.2856
en-fr	0.2987	0.2988
en-ru	0.1582	
cs-en	0.2328	0.2367
de-en	0.2389	0.2436
es-en	0.2916	0.2975
fr-en	0.2887	
ru-en	0.1975	0.2003
cs-de	0.1595	
cs-es	0.2170	0.2220
cs-fr	0.2220	0.2196
cs-ru	0.1660	
de-cs	0.1488	
es-cs	0.1580	
fr-cs	0.1420	
ru-cs	0.1506	

Table 3: BLEU scores of the baseline experiments (left column) on News Test 2013 data, computed by the system on tokenized data, versus similar setup with Gigaword. The difference was typically not significant.

3.2.3 Larger Parallel Data

Various combinations with larger parallel corpora were also tested. We do not have results for all combinations because these experiments needed a lot of time and resources and not all of them finished in time successfully.

In general the UN corpus seems to be of low quality or too much off-domain. It may help a little if used in combination with news-euro. If used separately, it always hurts the results.

The Giga French-English corpus gave the best results for English-French as expected, even without the core news-euro data. However, training the model on data of this size is extremely demanding on memory and time.

Finally, Czeng undoubtedly improves Czech-English translation in both directions. The news-euro dataset is smaller for this language pair, which makes Czeng stand out even more. See Table 4 for details.

3.2.4 Final Results

Table 5 compares our BLEU scores with those computed at `matrix.statmt.org`.

BLEU (without flag) denotes BLEU score computed by our system, comparing truecased tokenized hypothesis with truecased tokenized reference translation.

The official evaluation by `matrix.statmt.org` gives typically lower numbers, reflecting the loss caused by detokenization and new (different) tokenization.

Dir	Parallel	Mono	*BLEU*
en-es	news-euro	+gigaword	0.2856
en-es	news-euro-un	+gigaword	0.2844
en-es	un	un+gigaw.	0.2016
en-fr	giga	+gigaword	0.3106
en-fr	giga	+newsall	0.3037
en-fr	news-euro-un	+gigaword	0.3010
en-fr	news-euro	+gigaword	0.2988
en-fr	un	un	0.2933
es-en	news-euro	+gigaword	0.2975
es-en	news-euro-un	baseline	0.2845
es-en	un	un+news	0.2067
fr-en	news-euro-un	+gigaword	0.2914
fr-en	news-euro	baseline	0.2887
fr-en	un	un+news	0.2737

Table 4: BLEU scores with different parallel corpora.

3.2.5 Efficiency

The baseline experiments were conducted mostly on 64bit AMD Opteron quad-core 2.8 GHz CPUs with 32 GB RAM (decoding run on 15 machines in parallel) and the whole pipeline typically required between a half and a whole day.

However, we used machines with up to 500 GB RAM to train the large language models and translation models. Aligning the UN corpora with Giza++ took around 5 days. Giga French-English corpus was even worse and required several weeks to complete. Using such a large corpus without pruning is not practical.

4 Extra Experiments with Russian

In a separate set of experiments, we tried to take a basic Moses framework and change the setup a little for better results on morphologically rich languages.

Tried combinations were Russian-Czech and Russian-English.

Direction	*BLEU*	$BLEU_l$	$BLEU_t$
en-cs	0.1786	0.180	0.170
en-de	0.1833	0.179	0.173
en-es	0.2856	0.288	0.271
en-fr	0.3010	0.270	0.259
en-ru	0.1582	0.142	0.142
cs-en	0.2527	0.259	0.244
de-en	0.2389	0.244	0.230
es-en	0.2856	0.288	0.271
fr-en	0.2887	0.294	0.280
ru-en	0.1975	0.203	0.191
cs-de	0.1595	0.159	0.151
cs-es	0.2220	0.225	0.210
cs-fr	0.2220	0.191	0.181
cs-ru	0.1660	0.150	0.149
de-cs	0.1488	0.151	0.142
es-cs	0.1580	0.160	0.152
fr-cs	0.1420	0.145	0.137
ru-cs	0.1506	0.151	0.144

Table 5: Final BLEU scores. *BLEU* is truecased computed by the system, $BLEU_l$ is the official lowercased evaluation by `matrix.statmt.org`. $BLEU_t$ is official truecased evaluation. Although lower official scores are expected, notice the larger gap in en-fr and cs-fr translation. There seems to be a problem in our French detokenization procedure.

4.1 Data

For the additional Russian-to-Czech systems, we used following parallel data:

- UMC 0.1 (Klyueva and Bojar, 2008) – tri-parallel set, consisting of news articles – 93,432 sentences
- data mined from movie subtitles (described in further detail below) – 2,324,373 sentences
- Czech-Russian part of InterCorp – a corpus from translation of fiction books (Čermák and Rosen, 2012) – 148,847 sentences

For Russian-to-English translation, we used combination of

- UMC 0.1 – 95,540 sentences
- subtitles – 1,790,209 sentences
- Yandex English-Russian parallel corpus [6] – 1,000,000 sentences
- wiki headlines from WMT website [7] – 514,859 sentences
- common crawl from WMT website – 878,386 sentences

Added together, Russian-Czech parallel data consisted of 2,566,615 sentences and English-Czech parallel data consisted of 4,275,961 sentences [8].

We also used 765 sentences from UMC003 as a devset for MERT training.

We used the following monolingual corpora to train language models. Russian:

- Russian sides of all the parallel data – 4,275,961 sentences
- News commentary from WMT website – 150,217 sentences
- News crawl 2012 – 9,789,861 sentences

For Czech:

- Czech sides of all the parallel data – 2,566,615 sentences
- Data downloaded from Czech news articles[9] – 1,531,403 sentences
- WebColl (Spoustová et al., 2010) – 4,053,223 sentences
- PDT [10] – 115,844 sentences
- Complete Czech Wikipedia – 3,695,172 sentences
- Sentences scraped from Czech social server okoun.cz – 580,249 sentences

For English:

- English sides of all the paralel data – 4,275,961 sentences
- News commentary from WMT website – 150,217 sentences

Table 6 and Table 7 shows the sizes of the training data.

[6] `https://translate.yandex.ru/corpus?lang=en`
[7] `http://www.statmt.org/wmt13/translation-task.html`
[8] some sentences had to be removed for technical reasons
[9] `http://thepiratebay.sx/torrent/7121533/`
[10] `http://ufal.mff.cuni.cz/pdt2.0/`

Corpus	SentPairs	Tok lng1	Tok lng2
cs-ru	2,566,615	19,680,239	20,031,688
en-ru	4,275,961	64,619,964	58,671,725

Table 6: Number of sentence pairs and tokens for every language pair.

Corpus	Sentences	Tokens
en mono	13,426,211	278,199,832
ru mono	13,701,213	231,076,387
cs mono	12,542,506	202,510,993

Table 7: Number of sentences and tokens for every language.

4.1.1 Tokenization, tagging

Czech and English data was tokenized and tagged using Morče tagger; Russian was tokenized and tagged using TreeTagger. TreeTagger also does lemmatization; however, we didn't use lemmas for alignment or translation models, since our experiments showed that primitive stemming got better results.

However, what is important to mention is that TreeTagger had problems with some corpora, mostly Common Crawl. For some reason, Russian TreeTagger has problems with "dirty" data—sentences in English, French or random non-unicode noise. It either slows down significantly or stops working at all. For this reason, we wrapped TreeTagger in a script that detected those hangs and replaced the erroneous Russian sentences with bogus, one-letter Russian sentences (we can't delete those, since the lines already exist in the opposite languages; but since the pair doesn't really make sense in the first place, it doesn't matter as much).

All the data are lowercased for all the models and we recase the letters only at the very end.

4.1.2 Subtitle data

For an unrelated project dealing with movie subtitles translation, we obtained data from OpenSubtitles.org for Czech and English subtitles. However, those data were not aligned on sentence level and were less structured—we had thousands of `.srt` files with some sort of metadata.

When exploiting the data from the subtitles, we made several observations:

- language used in subtitles is very different from the language used in news articles
- one of the easiest and most accurate sentence alignments in movie subtitles is the one based purely on the time stamps
- allowing bigger differences in the time stamps in the alignment produced more data, but less accurate
- the subtitles are terribly out of domain (as experiments with using *only* the subtitle data showed us), but adding the corpus mined from the subtitles *still* increases the accuracy of the translation
- allowing bigger differences in the time stamps and, therefore, more (albeit less accurate) data always led to better results in our tests.

In the end, we decided to pair as much subtitles as possible, even with the risk of some being misaligned, because we found out that this helped the most.

4.2 Translation model, language model

For alignment, we used primitive stemming that takes just first 6 letters from a word. We found out that using this "brute force" stemming—for reasons that will have to be explored in a further research—return better results than regular lemmatization, for both alignment and translation model, as described further.

For each language pair, we used a translation model with two translation tables, one of them as backoff model. More exactly, the primary translation is from a form to a combination of (lower case) form and tag, and the secondary backoff translation is from a "stem" described above to a combination of (lower case) form and tag.

We built two language models—one for tags and one for lower case forms.

The models were actually a mixed model using interpolate option in SRILM—we trained a different language model for each corpus, and then we mixed the language models using a small development set from UMC003.

4.3 Final Results

The final results from `matrix.statmt.org` are in the table Table 8. You might notice a sharp difference between lowercased and truecased BLEU—that is due to a technical error that we didn't notice before the deadline.

Direction	$BLEU_l$	$BLEU_t$
ru-cs	0.158	0.135
cs-ru	0.165	0.162
ru-en	0.224	0.174
en-ru	0.163	0.160

Table 8: Lowercased and cased BLEU scores

5 Conclusion

We have described two independent Moses-based SMT systems we used for the WMT 2013 shared task. We discussed experiments with large data for many language pairs from the point of view of both the translation accuracy and efficiency.

Acknowledgements

The work on this project was supported by the grant P406/11/1499 of the Czech Science Foundation (GAČR), and by the grant 639012 of the Grant Agency of Charles University (GAUK).

References

František Čermák and Alexandr Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3):411–427.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. In *Technical report TR-10-98, Computer Science Group*, Harvard, MA, USA, August. Harvard University.

Natalia Klyueva and Ondřej Bojar. 2008. UMC 0.1: Czech-Russian-english multilingual corpus. In *International Conference Corpus Linguistics*.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–184, Los Alamitos, California, USA. IEEE Computer Society Press.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Praha, Czechia, June. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Johanka Spoustová, Miroslav Spousta, and Pavel Pecina. 2010. Building a web corpus of czech. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, Denver, Colorado, USA.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, OH, USA. Association for Computational Linguistics.

Daniel Zeman. 2012. Data issues of the multilingual translation matrix. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 395–400, Montréal, Canada. Association for Computational Linguistics.