Data Issues of the Multilingual Translation Matrix

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

Abstract

We describe our experiments with phrase-based machine translation for the WMT 2012 Shared Task. We trained one system for 14 translation directions between English or Czech on one side and English, Czech, German, Spanish or French on the other side. We describe a set of results with different training data sizes and subsets.

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.

Our goal is to run one system under as similar conditions as possible to all fourteen translation directions, to compare their translation accuracies and see why some directions are easier than others. Future work will benefit from knowing what are the special processing needs for a given language pair. 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 mentioned above.

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 395–400,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

2 The Translation System

Our 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 Data and Pre-processing Pipeline

We applied our system to all the eight official language pairs. In addition, we also experimented with translation between Czech on one side and German, Spanish or French 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 v7 and Europarl v7 corpora.[4] Table 1 shows the sizes of the training data.

The News Test 2010 data set[5] (2489 sentences in each language) was used as development data for MERT. BLEU scores reported in this paper were computed on the News Test 2012 set (3003 sentences each language). We do not use the News Tests 2008, 2009 and 2011.

Corpus	SentPairs	Tokens lng1	Tokens lng2
cs-en	782,756	17,997,673	20,964,639
de-en	2,079,049	55,143,719	57,741,141
es-en	2,123,036	61,784,972	59,217,471
fr-en	2,144,820	69,568,241	59,939,548
de-cs	652,193	17,422,620	15,383,601
es-cs	692,118	20,189,811	16,324,910
fr-cs	686,300	22,220,780	16,190,365

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. Every line corresponds to the respective version of EuroParl + News Commentary.

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 and normalize the directed (opening/closing) quotation marks (i.e. `"quoted"` → “quoted”). The motivation is twofold here: First, we hope that paired quotation marks could occasionally work as brackets and better denote parallel phrases for Moses; second, if Moses learns to output directed quotation marks, subsequent detokenization will be easier.

The data are then tagged and lemmatized. We used the Morče tagger for Czech and English lemmatization and TreeTagger for German, Spanish and French 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.

[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/wmt12/translation-task.html\#download`

[5] `http://www.statmt.org/wmt12/translation-task.html`

3.1 Quotation Marks

A broad range of characters is used to represent quotation marks in the training data: straight ASCII quotation mark; Unicode directed quotation marks (U+2018 to U+201F); acute and grave accents; math symbols such as prime and double prime (U+2032 to U+2037) etc. Spaces around quotes in the original untokenized text ought to provide hints as to the direction of the quotes (no space between the opening quote and the next word, and no space between the closing quote and the previous word) but unfortunately there are numerous cases where superfluous spaces are inserted or required spaces are missing.

Nested quoting is also possible, such as in

As the Wise Men ’ s Report also says , and I quote : ’ It is elementary ’ common sense ’ that the Commission should have supported the Parliament ’ s decision - making process . ’

We want all possible quotation marks converted to one pair of characters. We do not mind the distinction between single and double quotes but we want to keep (or restore) the distinction between opening and closing quotes. In addition, we want to identify the apostrophe acting as grapheme in some languages, and keep it (or normalize it, as it could also be mis-typed as acute accent or something else):

As the Wise Men ’ s Report also says , and I quote : “ It is elementary “ common sense ” that the Commission should have supported the Parliament ’ s decision - making process . ”

We attempt at solving the problem by a set of rules that consider mutual positions of quotation marks, spaces and other punctuation, and also some language-dependent rules (especially on the lexical apostrophe, e.g. in French *d’, l’*).

Our rules applied to 1.84 % of Spanish sentences, 2.47 % Czech, 2.77 % German, 4.33 % English and 16.9 % French (measured on Europarl data).

Our approach is different from the normalization script provided and applied by the organizers of the shared task, which merely converts all quotes to the undirected ASCII characters. We believe that such MT output is incorrect, so we submitted two versions of each system run: the *primary* version is intended for human evaluation and does not apply the “official” normalization of punctuation. In contrast, the *secondary* version is normalized, which naturally leads to higher scores in the automatic evaluation.

4 Experiments

In the following section we describe several different settings and corpora combinations we experimented with. BLEU scores have been computed by our system, comparing truecased tokenized hypothesis with truecased tokenized reference translation.

Such scores must differ from the official evaluation—see Section 4.4 for discussion of the final results.

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

4.1 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).

4.2 Larger Monolingual Data

Besides the monolingual halves of the parallel corpora, additional monolingual data were provided / permitted:

- The Crawled News corpus from the years 2007 to 2011, various sizes for each language and year.
- The Gigaword corpora published by the Linguistic Data Consortium, available only for English (4^{th} edition), Spanish (3^{rd}) and French (3^{rd}).

Due to bugs in the lemmatizers, we were not able to process certain parts of the large corpora in time. Table 2 gives the sizes of the subsets available for our experiments and Table 3 compares BLEU scores with large language models against the baseline.

Corpus	Segments	Tokens
newsc+euro.cs	819,434	18,491,692
newsc+euro.de	2,360,811	58,683,607
newsc+euro.en	2,430,718	65,934,441
newsc+euro.es	2,307,429	66,072,443
newsc+euro.fr	2,361,764	74,083,166
news.all.cs	14,552,899	244,728,011
news.all.de	24,446,319	462,924,303
news.all.en	42,161,804	1,039,806,242
news.all.es	8,627,438	249,022,213
news.all.fr	16,708,622	438,489,352
gigaword.en	70,592,779	2,546,581,646
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.

The Crawled News corpora, in-domain and larger than the parallel corpora by an order of magnitude, turned out to help significantly improve the scores of all language pairs. On the other hand, and to our surprise, we were not able to achieve any further improvement by using the Gigaword corpora. Taking into account the extra requirements on memory when building such big language models, this makes the usefulness of Gigaword questionable. We have no plausible explanation at the moment.

4.3 Larger Parallel Data

Even stranger behavior was observed when adding the large UN parallel corpus (over 10 million sentence pairs). When used separately (even for language model) it decreased BLEU significantly, which could be explained by different domain. When used together with News Commentary and Europarl, and with a language model trained on the Crawled News corpus, it barely outperformed the same setting without the UN corpus.[6] However, the es-en direction is a notable exception where the UN corpus alone gave by far the best score. See Table 4 for details.

Direction	Baseline	news.all	gigaword
en-cs	0.1196	0.1434	
en-de	0.1426	0.1629	
en-es	0.2778	0.3136	0.3136
en-fr	0.2599	0.2897	0.2874
cs-en	0.1796	0.2031	0.2013
de-en	0.1877	0.2136	0.2144
es-en	0.2219	0.2428	0.2390
fr-en	0.2459	0.2764	0.2756
cs-de	0.1365	0.1550	
cs-es	0.1952	0.2211	0.2184
cs-fr	0.1953	0.2167	0.2147
de-cs	0.1212	0.1400	
es-cs	0.1281	0.1489	
fr-cs	0.1253	0.1442	

Table 3: BLEU scores of the baseline experiments (left column) on News Test 2012 data, computed by the system on tokenized data, versus similar setup with large monolingual corpus (news.all, middle column). Gigaword never brought significant improvement.

We failed to lemmatize the giga French-English corpus in time, so we do not present any results with that corpus.

4.4 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.

[6] One of the anonymous reviewers mentioned that the quality of the UN corpus is relatively low. That could explain our observations.

Direction	Parallel	Mono	*BLEU*
en-es	news-euro-un	news.all	0.3194
en-es	news-euro	news.all	0.3136
en-es	un	un	0.2694
en-fr	news-euro	news.all	0.2897
en-fr	un	un	0.2541
es-en	un	un	**0.2688**
es-en	news-euro	news.all	0.2428
fr-en	news-euro	news.all	0.2764
fr-en	un	un	0.2392

Table 4: BLEU scores with different parallel corpora.

4.5 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.

5 Conclusion

We have described the Moses-based SMT system we used for the WMT 2012 shared task. We discussed experiments with large data for many language pairs from the point of view of both the translation accuracy and efficiency. We were unable to process all data that was available; even the experiments where we did use larger data did not outperform the smaller experiments significantly. Nevertheless, using the Crawled News monolingual corpus proved essential.

Acknowledgements

The work on this project was supported by the grant P406/11/1499 of the Czech Science Foundation (GAČR).

Direction	$BLEU$	$BLEU_l$	$BLEU_t$
en-cs	0.1434	0.144	0.136
en-de	0.1629	0.159	0.154
en-es	0.3136	0.316	0.297
en-fr	0.2897	0.263	0.251
cs-en	0.2031	0.207	0.192
de-en	0.2136	0.214	0.200
es-en	0.2428	0.253	0.240
fr-en	0.2764	0.280	0.266
cs-de	0.1550	0.153	0.147
cs-es	0.2211	0.224	0.207
cs-fr	0.2167	0.197	0.186
de-cs	0.1400	0.141	0.134
es-cs	0.1489	0.150	0.143
fr-cs	0.1442	0.145	0.138

Table 5: BLEU scores with the large language models. $BLEU$ is 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.

References

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.

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.

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.