Hierarchical Phrase-Based MT at the Charles University for the WMT 2010 Shared Task

Daniel Zeman
Charles University in Prague, Institute of Formal and Applied Linguistics (ÚFAL)
Univerzita Karlova v Praze, Ústav formální a aplikované lingvistiky (ÚFAL)
Malostranské náměstí 25, Praha, CZ-11800, Czechia
zeman@ufal.mff.cuni.cz

Abstract

We describe our experiments with hierarchical phrase-based machine translation for WMT 2010 Shared Task. We provide a detailed description of our configuration and data so the results are replicable. For English-to-Czech translation, we experiment with several datasets of various sizes and with various preprocessing sequences. For the other 7 translation directions, we just present the baseline results.

1 Introduction

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.

There are numerous ways how these issues could be addressed. For instance, parsing and syntax-aware reordering of the source-language sentences can help with the word order differences (same goal could be achieved by a reordering model or a synchronous context-free grammar in a hierarchical system). Factored translation, a secondary language model of morphological tags or even a morphological generator are some of the possible solutions to the poor-to-rich translation issues.

Our submission to the shared task should reveal where a pure hierarchical system stands in this jungle and what of the above mentioned ideas match the phenomena the system suffers from. Although our primary focus lies on English-to-Czech translation, we also report the accuracy of the same system on moderately-sized corpora for the other three languages and seven translation directions.

2 The Translation System

Our translation system belongs to the hierarchical phrase-based class (Chiang, 2007), i.e. phrase pairs with nonterminals (rules of a synchronous context-free grammar) are extracted from symmetrized word alignments and subsequently used by the decoder. We use Joshua, a Java-based open-source implementation of the hierarchical decoder (Li et al., 2009), release 1.1.[1]

Word alignment was computed using the first three steps of the `train-factored-phrase-model.perl` script packed with Moses[2] (Koehn et al., 2007). This includes the usual combination of word clustering using `mkcls`[3] (Och, 1999), two-way word alignment using GIZA++[4] (Och and Ney, 2003), and alignment symmetrization using the *grow-diag-final-and* heuristic (Koehn et al., 2003).

For language modeling we use the SRILM toolkit[5] (Stolcke, 2002) with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998).

We use the Z-MERT implementation of minimum error rate training (Zaidan, 2009). The following settings have been used for Joshua and Z-MERT:

[1] `http://sourceforge.net/projects/joshua/`
[2] `http://www.statmt.org/moses/`
[3] `http://fjoch.com/mkcls.html`
[4] `http://fjoch.com/GIZA++.html`
[5] `http://www-speech.sri.com/projects/srilm/`

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 212–215,
Uppsala, Sweden, 15-16 July 2010.

- Grammar extraction: `--maxPhraseLength=5`
- Decoding: `span_limit=10 fuzz1=0.1 fuzz2=0.1 max_n_items=30 relative_threshold=10.0 max_n_rules=50 rule_relative_threshold=10.0`
- N-best decoding: `use_unique_nbest=true use_tree_nbest=false add_combined_cost=true top_n=300`
- Z-MERT: `-m BLEU 4 closest -maxIt 5 -ipi 20`

3 Data and Pre-processing Pipeline

3.1 Baseline Experiments

We applied our system to all eight language pairs. However, for all but one we ran only a baseline experiment. From the data point of view the baseline experiments were even more constrained than the organizers of the shared task suggested. We did not use the Europarl corpus, we only used the News Commentary corpus[6] for training. The target side of the News Commentary corpus was also the only source to train the language model. Table 1 shows the size of the corpus.

Corpus	SentPairs	Tokens xx	Tokens en
cs-en	94,742	2,077,947	2,327,656
de-en	100,269	2,524,909	2,484,445
es-en	98,598	2,742,935	2,472,860
fr-en	84,624	2,595,165	2,137,407

Table 1: Number of sentence pairs and tokens for every language pair in the News Commentary corpus. Unlike the organizers of the shared task, we stick with the standard ISO 639 language codes: cs = Czech, de = German, en = English, es = Spanish, fr = French.

Note that in some cases the grammar extraction algorithm in Joshua fails if the training corpus contains sentences that are too long. Removing sentences of 100 or more tokens (per advice by Joshua developers) effectively healed all failures. Unfortunately, for the baseline corpora the loss of training material was still considerable and resulted in drop of BLEU score, though usually insignificant.[7]

[6] Available for download at `http://www.statmt.org/wmt10/translation-task.html` using the link "Parallel corpus training data".

[7] Table 1 and Table 2 present statistics *before* removing the long sentences.

The News Test 2008 data set (2051 sentences in each language) was used as development data for MERT. BLEU scores reported in this paper were computed on the News Test 2009 set (2525 sentences each language). The official scores on News Test 2010 are given only in the main WMT 2010 paper.

Only lowercased data were used for the baseline experiments.

3.2 English-to-Czech

A separate set of experiments has been conducted for the English-to-Czech direction and larger data were used. We used CzEng 0.9 (Bojar and Žabokrtský, 2009)[8] as our main parallel corpus. Following CzEng authors' request, we did not use sections 8* and 9* reserved for evaluation purposes.

As the baseline training dataset ("Small" in the following) only the news section of CzEng was used. For large-scale experiments ("Large" in the following), we used all CzEng together with the EMEA corpus[9] (Tiedemann, 2009).[10]

As our monolingual data we use the monolingual data provided by WMT10 organizers for Czech. Table 2 shows the sizes of these corpora.

Corpus	SentPairs	Tokens cs	Tokens en
Small	126,144	2,645,665	2,883,893
Large	7,543,152	79,057,403	89,018,033
Mono	13,042,040	210,507,305	

Table 2: Number of sentences and tokens in the Czech-English corpora.

Again, the official WMT 2010[11] development set (News Test 2008, 2051 sentences each language) and test set (News Test 2009, 2525 sentences each language) are used for MERT and evaluation, respectively. The official scores on News Test 2010 are given only in the main WMT 2010 paper.

We use a slightly modified tokenization rules compared to CzEng export format. Most notably, we normalize English abbreviated negation and auxiliary verbs ("couldn't" → "could not") and

[8] `http://ufal.mff.cuni.cz/czeng/`

[9] `http://urd.let.rug.nl/tiedeman/OPUS/EMEA.php`

[10] Unfortunately, the EMEA corpus is badly tokenized on the Czech side with fractional numbers split into several tokens (e.g. "3, 14"). We attempted to reconstruct the original detokenized form using a small set of regular expressions.

[11] `http://www.statmt.org/wmt10`

attempt at normalizing quotation marks to distinguish between opening and closing one following proper typesetting rules.

The rest of our pre-processing pipeline matches the processing employed in CzEng (Bojar and Žabokrtský, 2009).[12] We use "supervised truecasing", meaning that 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.

4 Experiments

All BLEU scores were computed directly by Joshua on the News Test 2009 set. Note that they differ from what the official evaluation script would report, due to different tokenization.

4.1 Baseline Experiments

The set of baseline experiments with all translation directions involved running the system on lowercased News Commentary corpora. Word alignments were computed on 4-character stems (including the en-cs and cs-en directions). A trigram language model was trained on the target side of the parallel corpus.

Direction	BLEU
en-cs	0.0905
en-de	0.1114
cs-en	0.1471
de-en	0.1617
en-es	0.1966
en-fr	0.2001
fr-en	0.2020
es-en	0.2025

Table 3: Lowercased BLEU scores of the baseline experiments on News Test 2009 data.

4.2 English-to-Czech

The extended (non-baseline) English-to-Czech experiments were trained on larger parallel and monolingual data, described in Section 3.2. Note that the dataset denoted as "Small" still falls into the constrained task because it only uses CzEng 0.9 and the WMT 2010 monolingual data.

[12] Due to the subsequent processing, incl. parsing, the tokenization of English follows PennTreebenk style. The rather unfortunate convention of treating hyphenated words as single tokens increases our out-of-vocabulary rate.

Word alignments were computed on lemmatized version of the parallel corpus. Hexagram language model was trained on the monolingual data. Truecased data were used for training, as described above; the BLEU scores of these experiments in Table 4 are computed on truecased system output.

Setup	BLEU
Baseline	0.0905
Small	0.1012
Large	0.1300

Table 4: BLEU scores (lowercased baseline, truecased rest) of the English-to-Czech experiments, including the baseline experiment with News Commentary, mentioned earlier.

As for the official evaluation on News Test 2010, we used the Small setup as our *primary submission*, and the Large setup as *secondary* despite its better results. The reason was that it was not clear whether the experiment would be finished in time for the official evaluation.[13]

An interesting perspective on the three en-cs models is provided by the feature weights optimized during MERT. We can see in Table 5 that the small and relatively weak baseline LM is trusted less than the most influential translation feature while for large parallel data and even much larger LM the weights are distributed more evenly.

Setup	LM	Pt_0	Pt_1	Pt_2	WP
Baseline	1.0	1.55	0.51	0.63	−2.63
Small	1.0	1.03	0.72	−0.09	−0.34
Large	1.0	0.98	0.97	−0.02	−0.82

Table 5: Feature weights are relative to the weight of LM, the score by the language model. Then there are the three translation features: $Pt_0 = P(e|f)$, $Pt_1 = P_{lex}(f|e)$ and $Pt_2 = P_{lex}(e|f)$. WP is the word penalty.

4.3 Efficiency

The machines on which the experiments were conducted are 64bit Intel Xeon dual core 2.8 GHz CPUs with 32 GB RAM.

Word alignment of each baseline corpus took about 1 hour, time needed for data preprocessing

[13] In fact, it was not finished in time. Due to a failure of a MERT run, we used feature weights from the primary submission for the secondary one, too.

and training of the language model was negligible. Grammar extraction took about four hours but it could be parallelized. For decoding the test data were split into 20 chunks that were processed in parallel. One MERT iteration, including decoding, took from 30 minutes to 1 hour.

Training the large en-cs models requires more careful engineering. The grammar extraction easily consumes over 20 GB memory so it is important to make sure Java really has access to it. We parallelized the extraction in the same way as we had done with the decoding; even so, about 5 hours were needed to complete the extraction. The decoder now must use the SWIG-linked SRILM library because Java-based language modeling is too slow and memory-consuming. Otherwise, the decoding times are comparable to the baseline experiments.

5 Conclusion

We have described the hierarchical phrase-based SMT system we used for the WMT 2010 shared task. For English-to-Czech translation, we discussed experiments with large data from the point of view of both the translation accuracy and efficiency.

This has been our first attempt to switch to hierarchical SMT and we have not gone too far beyond just putting together the infrastructure and applying it to the available data. Nevertheless, our en-cs experiments not only confirm that more data helps; in the Small and Large setup, the data was not only larger than in Baseline, it also underwent a more refined preprocessing. In particular, we took advantage of the Czeng corpus being lemmatized to produce better word alignment; also, the truecasing technique helped to better target named entities.

Acknowledgements

The work on this project was supported by the grant MSM0021620838 by the Czech Ministry of Education.

References

Ondřej Bojar and Zdeněk Žabokrtský. 2009. Czeng 0.9: Large parallel treebank with rich annotation. *The Prague Bulletin of Mathematical Linguistics*, 92:63–83.

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.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228.

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.

Zhifei Li, Chris Callison-Burch, Sanjeev Khudanpur, and Wren Thornton. 2009. Decoding in Joshua: Open Source, Parsing-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 91:47–56, 1.

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

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 71–76, Bergen, Norway, June. 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.

Jörg Tiedemann. 2009. News from opus – a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing (vol. V)*, pages 237–248. John Benjamins.

Omar F. Zaidan. 2009. Z-mert: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.