Hierarchical Phrase-Based MT at the Charles University for the WMT 2011 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 the WMT 2011 Shared Task. We trained a system for all 8 translation directions between English on one side and Czech, German, Spanish or French on the other side, though we focused slightly more on the English-to-Czech direction. We provide a detailed description of our configuration and data so the results are replicable.

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 the following 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.

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 goal is to run one system under as similar conditions as possible to all eight 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. Still, comparability of the results is limited, as the quality and quantity of English-Czech data differs from that of the other pairs.

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 496–500,
Edinburgh, Scotland, UK, July 30–31, 2011.

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 1.3.[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 (for the sake of reproducibility, we keep the original names of the options; for their detailed explanation please refer to the documentation available on-line at the Joshua project site). `-ipi` is the number of intermediate initial points per Z-MERT iteration.

- Grammar extraction: `maxPhraseSpan=10 maxPhraseLength=5 maxNonterminals=2 maxNonterminalSpan=2 requireTightSpans=true edgeXViolates=true sentenceInitialX=true sentenceFinalX=true ruleSampleSize=300`
- Language model order: 6 (hexagram)
- 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

We applied our system to all eight language pairs. From the data point of view the experiments were even more constrained than the organizers of the shared task suggested. We used neither the French/Spanish-English UN corpora nor the 10^9 French-English corpus. For 7 translation directions we used the Europarl ver6 and News-Commentary ver6 corpora[6] for training. The target side of the corpora was our only source of monolingual data for training the language model. Table 1 shows the size of the training data.

For the English-Czech direction, we used CzEng 0.9 (Bojar and Žabokrtský, 2009)[7] as our main parallel corpus. Following CzEng authors' request, we did not use sections 8* and 9* reserved for evaluation purposes.

In addition, we also used the EMEA corpus[8] (Tiedemann, 2009).[9]

Czech was also the only language where we used extra monolingual data for the language model. It was the set provided by the organizers of WMT 2010 (13,042,040 sentences, 210,507,305 tokens).

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 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).[10] 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

[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/`

[6] Available for download at `http://www.statmt.org/wmt11/translation-task.html` using the link "Parallel corpus training data".
[7] `http://ufal.mff.cuni.cz/czeng/`
[8] `http://urd.let.rug.nl/tiedeman/OPUS/EMEA.php`
[9] 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.

Corpus	SentPairs	Tokens xx	Tokens en
cs-en	583,124	13,224,596	15,397,742
de-en	1,857,087	48,834,569	51,243,594
es-en	1,903,562	54,488,621	52,369,658
fr-en	1,920,363	61,030,918	52,686,784
en-cs	7,543,152	79,057,403	89,018,033

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. The en-cs line describes the CzEng + EMEA combined corpus, all other lines correspond to the respective versions of EuroParl + News Commentary.

words are lowercased.

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.[11]

The News Test 2008 data set[12] (2051 sentences in each language) was used as development data for MERT. BLEU scores reported in this paper were computed on the News Test 2011 set (3003 sentences each language). We do not use the News Test 2009 and 2010.

4 Experiments

All BLEU scores were computed directly by Joshua on the News Test 2011 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 lowercased 4-character stems. A hexagram language model was trained on the target side of the parallel corpus.

In the en-cs case, 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 score of this experiment in Table 2 is computed on truecased system output.

Direction	$BLEU_J$	$BLEU_l$	$BLEU_t$
en-cs	0.1274	0.141	0.123
en-de	0.1324	0.128	0.052
en-es	0.2756	0.274	0.221
en-fr	0.2727	0.212	0.174
cs-en	0.1782	0.178	0.137
de-en	0.1957	0.187	0.137
es-en	0.2630	0.255	0.197
fr-en	0.2471	0.248	0.193

Table 2: Lowercased BLEU scores of the baseline experiments on News Test 2011 data: $BLEU_J$ is computed by the system, $BLEU_l$ is the official evaluation by `matrix.statmt.org` (it differs because of different tokenization). $BLEU_t$ is official truecased evaluation.

An interesting perspective on the models is provided by the feature weights optimized during MERT. We can see in Table 3 that translation models are trusted significantly more than language models for the en-de, de-en and es-en directions. In fact, the language model has a low relative weight in all language pairs but en-cs, which was the only pair where we used a significant amount of extra monolingual data. In the future, we should probably use the Gigaword corpus for the to-English directions.

Setup	LM	Pt_0	Pt_1	Pt_2	WP
en-cs	1.0	1.04	0.84	−0.06	−1.19
en-de	1.0	2.60	0.57	0.47	−3.17
en-es	1.0	1.67	0.81	0.60	−2.96
en-fr	1.0	1.41	0.92	0.53	−2.80
cs-en	1.0	1.48	0.94	1.08	−4.55
de-en	1.0	2.28	1.11	0.34	−2.88
es-en	1.0	2.26	1.67	0.23	−0.84
fr-en	1.0	1.89	1.32	0.13	−0.04

Table 3: 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.

[10] 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.

[11] Table 1 presents statistics *before* removing the long sentences.

[12] `http://www.statmt.org/wmt11/translation-task.html`

4.2 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 parallel corpus was the most resource-consuming subtask. It took between 12 and 48 hours, though it could be cut to one half by running both GIZA++ directions in parallel. The time needed for data preprocessing and training of the language model was negligible. Parallelized grammar extraction took 19 processors for about an hour. 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 of large models requires some careful engineering. The grammar extraction easily consumes over 20 GB memory so it is important to make sure Java really has access to it. The decoder must use the SWIG-linked SRILM library because Java-based language modeling is too slow and memory-consuming.

4.3 Supervised Truecasing

Our baseline experiments operated on lowercased data, except for en-cs, where truecased word forms were obtained using lemmas from morphological annotation (note that guessing of the true case is only needed for the sentence-initial token, other words can just be left in their original form).

As contrastive runs we applied the supervised truecasing to other directions as well. We used the Morče tagger for English lemmatization, TreeTagger for German and two simple rule-based approaches to Spanish and French lemmatization. All these tools are embedded in the TectoMT analysis framework (Žabokrtský et al., 2008).

The results are in Table 4. $BLEU_t$ has increased in all cases w.r.t. the baseline results.

4.4 Alignment on Lemmas

Once we are able to lemmatize all five languages we can also experiment with word alignments based on lemmas. Table 5 shows that the differences in BLEU are insignificant.

5 Conclusion

We have described the hierarchical phrase-based SMT system we used for the WMT 2011 shared task. We discussed experiments with large data from the point of view of both the translation accuracy and efficiency. We used moderately-sized training data and took advantage from their basic linguistic annotation (lemmas). The truecasing technique helped us to better target named entities.

Direction	$BLEU_J$	$BLEU_l$	$BLEU_t$
en-cs	0.1191	0.126	0.119
en-de	0.1337	0.131	0.127
en-es	0.2573	0.276	0.265
en-fr	0.2591	0.211	0.189
cs-en	0.1692	0.180	0.168
de-en	0.1885	0.191	0.178
es-en	0.2446	0.260	0.236
fr-en	0.2243	0.245	0.221

Table 4: Results of experiments with supervised truecasing. Note that training on truecased corpus slightly influenced even the lowercased BLEU (cf. with Table 2). This is because probabilities of tokens that may appear both uppercased and lowercased (with different meanings) have changed, and thus different translation may have been chosen.

Direction	$BLEU_Jl4$	$BLEU_Jlm$
en-cs	0.1191	0.1193
en-de	0.1337	0.1318
en-es	0.2573	0.2590
en-fr	0.2591	0.2592
cs-en	0.1692	0.1690
de-en	0.1885	0.1892
es-en	0.2446	0.2452
fr-en	0.2243	0.2244

Table 5: Results of experiments with word alignment computed on different factors. $BLEU_Jl4$ is the score computed by Joshua on lowercased test data for the original experiments (alignment based on lowercased 4-character prefixes). $BLEU_Jlm$ is the corresponding score for alignment based on lemmas.

Acknowledgements

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

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.

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.

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.