Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation *

Ondřej Bojar and **Jan Hajič**
Institute of Formal and Applied Linguistics
ÚFAL MFF UK, Malostranské náměstí 25
CZ-11800 Praha, Czech Republic
{bojar,hajic}@ufal.mff.cuni.cz

Abstract

This paper describes our two contributions to WMT08 shared task: factored phrase-based model using Moses and a probabilistic tree-transfer model at a deep syntactic layer.

1 Introduction

Czech is a Slavic language with very rich morphology and relatively free word order. The Czech morphological system (Hajič, 2004) defines 4,000 tags in theory and 2,000 were actually seen in a big tagged corpus while the English Penn Treebank tagset contains just about 50 tags. In our parallel corpus (see below), the English vocabulary size is 148k distinct word forms but more than twice as big in Czech, 343k distinct word forms.

When translating to Czech from an analytic language such as English, target word forms have to be chosen correctly to produce a grammatical sentence and preserve the expressed relations between elements in the sentence, e.g. verbs and their modifiers.

This year, we have taken two radically different approaches to English-to-Czech MT. Section 2 describes our setup of the phrase-based system Moses (Koehn et al., 2007) and Section 3 focuses on a system with probabilistic tree transfer employed at a deep syntactic layer and the new challenges this approach brings.

*The work on this project was supported by the grants FP6-IST-5-034291-STP (EuroMatrix), MSM0021620838, MŠMT ČR LC536, and GA405/06/0589.

2 Factored Phrase-Based MT to Czech

Bojar (2007) describes various experiments with factored translation to Czech aimed at improving target-side morphology. We use essentially the same setup with some cleanup and significantly larger target-side training data:

Parallel data from CzEng 0.7 (Bojar et al., 2008), with original sentence-level alignment and tokenization. The parallel corpus was taken as a monolithic text source disregarding differences between CzEng data sources. We use only 1-1 aligned sentences.

Word alignment using GIZA++ toolkit (Och and Ney, 2000), the default configuration as available in training scripts for Moses. We based the word alignment on Czech and English lemmas (base forms of words) as provided by the combination of taggers and lemmatizers by Hajič (2004) for Czech and Brants (2000) followed by Minnen et al. (2001) for English. We symmetrized the two GIZA++ runs using grow-diag-final heuristic.

Truecasing. We attempted to preserve meaning-bearing case distinctions. The Czech lemmatizer produces case-sensitive lemmas and thus makes it easy to cast the capitalization of the lemma back on the word form.[1] For English we approximate the same effect by a two-step procedure.[2]

[1]We change the capitalization of the form to match the lemma in cases where the lemma is lowercase, capitalized (ucfirst) or all-caps. For mixed-case lemmas, we keep the form intact.

[2]We first collect a lexicon of the most typical "shapes" for each word form (ignoring title-like sentences with most words capitalized and the first word in a sentence). Capitalized and all-caps words in title-like sentences are then changed to their

Proceedings of the Third Workshop on Statistical Machine Translation, pages 143–146,
Columbus, June 2008.

Decoding steps. We use a simple two-step scenario similar to class-based models (Brown and others, 1992): (1) the source English word forms are translated to Czech word forms and (2) full Czech morphological tags are generated from the Czech forms.

Language models. We use the following 6 independently weighted language models for the target (Czech) side:

- 3-grams of word forms based on all CzEng 0.7 data, 15M tokens,
- 3-grams of word forms in Project Syndicate section of CzEng (in-domain for WMT07 and WMT08 NC-test set), 1.8M tokens,
- 4-grams of word forms based on Czech National Corpus (Kocek et al., 2000), version SYN2006, 365M tokens,
- three models of 7-grams of morphological tags from the same sources.

Lexicalized reordering using the monotone/swap/discontinuous bidirectional model based on both source and target word forms.

MERT. We use the minimum-error rate training procedure by Och (2003) as implemented in the Moses toolkit to set the weights of the various translation and language models, optimizing for BLEU.

Final detokenization is a simple rule-based procedure based on Czech typographical conventions. Finally, we capitalize the beginnings of sentences.

See BLEU scores in Table 2 below.

3 MT with a Deep Syntactic Transfer

3.1 Theoretical Background

Czech has a well-established theory of linguistic analysis called Functional Generative Description (Sgall et al., 1986) supported by a big treebanking enterprise (Hajič and others, 2006) and on-going adaptations for other languages including English (Cinková and others, 2004). There are two layers of syntactic analysis, both formally captured as labelled ordered dependency trees: the ANALYTICAL (a-, surface syntax) representation bears a 1-1 correspondence between tokens in the sentence and nodes in the tree; the TECTOGRAMMATICAL (t-, deep syntax) representation contains nodes only for autosemantic words and adds nodes for elements not expressed on the surface but required by the grammar (e.g. dropped pronouns).

We use the following tools to automatically annotate plaintext up to the t-layer: (1) TextSeg (Češka, 2006) for tokenization, (2) tagging and lemmatization see above, (3) parsing to a-layer: Collins (1996) followed by head-selection rules for English, McDonald and others (2005) for Czech, (4) parsing to t-layer: Žabokrtský (2008) for English, Klimeš (2006) for Czech.

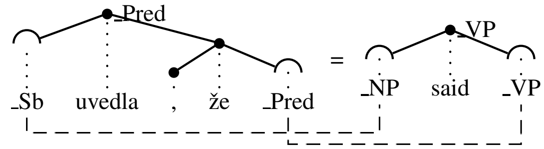

Figure 1: Sample treelet pair, a-layer.

3.2 Probabilistic Tree Transfer

The transfer step is based on Synchronous Tree Substitution Grammars (STSG), see Bojar and Čmejrek (2007) for a detailed explanation. The essence is a log-linear model to search for the most likely synchronous derivation $\hat{\delta}$ of the source T_1 and target T_2 dependency trees:

$$\hat{\delta} = \underset{\delta \text{ s.t. source is } T_1}{\operatorname{argmax}} \exp\Big(\sum_{m=1}^{M} \lambda_m h_m(\delta)\Big) \qquad (1)$$

The key feature function h_m in STSG represents the probability of attaching pairs of dependency treelets $t^i_{1:2}$ such as in Figure 1 into aligned pairs of frontiers ($\frown$) in another treelet pair $t^j_{1:2}$ given frontier state labels (e.g. _Pred-_VP in Figure 1):

$$h_{STSG}(\delta) = \log \prod_{i=0}^{k} p(t^i_{1:2} \mid \text{frontier states}) \qquad (2)$$

Other features include e.g. number of internal nodes (drawn as $\bullet$ in Figure 1) produced, number of treelets produced, and more importantly the traditional n-gram language model if the target (a-)tree

typical shape. In other sentences we change the case only if a typically lowercase word is capitalized (e.g. at the beginning of the sentence) or if a typically capitalized word is all-caps. Unknown words in title-like sentences are lowercased and left intact in other sentences.

is linearized right away or a binode model promoting likely combinations of the governor $g(e)$ and the child $c(e)$ of an edge $e \in T_2$:

$$h_{binode}(\delta) = \log \prod_{e \in T_2} p(c(e) \mid g(e)) \qquad (3)$$

The probabilistic dictionary of aligned treelet pairs is extracted from node-aligned (GIZA++ on linearized trees) parallel automatic treebank as in Moses' training: all treelet pairs compatible with the node alignment.

3.2.1 Factored Treelet Translation

Labels of nodes at the t-layer are not atomic but consist of more than 20 attributes representing various linguistic features.[3] We can consider the attributes as individual factors (Koehn and Hoang, 2007). This allows us to condition the translation choice on a subset of source factors only. In order to generate a value for each target-side factor, we use a sequence of mapping steps similar to Koehn and Hoang (2007). For technical reasons, our current implementation allows to generate factored target-side only when translating a single node to a single node, i.e. preserving the tree structure.

In our experiments we used 8 source (English) t-node attributes and 14 target (Czech) attributes.

3.3 Recent Experimental Results

Table 1 shows BLEU scores for various configurations of our decoder. The abbreviations indicate between which layers the tree transfer was employed (e.g. "eact" means English a-layer to Czech t-layer). The "p" layer is an approximation of phrase-based MT: the surface "syntactic" analysis is just a left-to-right linear tree.[4] For setups ending in t-layer, we use a deterministic generation the of Czech sentence by Ptáček and Žabokrtský (2006).

For WMT08 shared task, Table 2, we used a variant of the "etct factored" setup with the annotation pipeline as incorporated in TectoMT (Žabokrtský, 2008) environment and using TectoMT internal rules for t-layer parsing and generation instead of Klimeš (2006) and (Ptáček and Žabokrtský, 2006).

Tree-based Transfer	LM Type	BLEU
epcp	n-gram	10.9±0.6
eaca	n-gram	8.8±0.6
epcp	none	8.7±0.6
eaca	none	6.6±0.5
etca	n-gram	6.3±0.6
etct factored, preserving structure	binode	5.6±0.5
etct factored, preserving structure	none	5.3±0.5
eact, target side atomic	binode	3.0±0.3
etct, atomic, all attributes	binode	2.6±0.3
etct, atomic, all attributes	none	1.6±0.3
etct, atomic, just t-lemmas	none	0.7±0.2
Phrase-based (Moses) as reported by Bojar (2007)		
Vanilla	n-gram	12.9±0.6
Factored to improve target morphology	n-gram	14.2±0.7

Table 1: English-to-Czech BLEU scores for syntax-based MT on WMT07 DevTest.

	WMT07	WMT08	
	DevTest	NC Test	News Test
Moses	14.9±0.9	16.4±0.6	12.3±0.6
Moses, CzEng data only	13.9±0.9	15.2±0.6	10.0±0.5
etct, TectoMT annotation	4.7±0.5	4.9±0.3	3.3±0.3

Table 2: WMT08 shared task BLEU scores.

3.3.1 Discussion

Our syntax-based approach does not reach scores of phrase-based MT due to the following reasons:

Cumulation of errors at every step of analysis.

Data loss due to incompatible parses and node alignment. Unlike e.g. Quirk et al. (2005) or Huang et al. (2006) who parse only one side and project the structure, we parse both languages independently. Natural divergence and random errors in either of the parses and/or the alignment prevent us from extracting many treelet pairs.

Combinatorial explosion in target node attributes. Currently, treelet options are fully built in advance. Uncertainty in the many t-node attributes leads to too many insignificant variations while e.g. different lexical choices are pushed off the stack. While vital for final sentence generation (see Table 1), fine-grained t-node attributes should be produced only once all key structural, lexical and form decisions have been made. The same sort of explosion makes complicated factored setups not yet feasible in Moses, either.

[3] Treated as atomic, t-node labels have higher entropy (11.54) than lowercase plaintext (10.74). The t-layer by itself does not bring any reduction in vocabulary. The idea is that the attributes should be more or less independent and should map easier across languages.

[4] Unlike Moses, "epcp" does not permit phrase reordering.

Lack of n-gram LM in the (deterministic) generation procedures from a t-tree. While we support final LM-based rescoring, there is too little variance in n-best lists due to the explosion mentioned above.

Too many model parameters given our stack limit. We use identical MERT implementation to optimize λ_ms but in the large space of hypotheses, MERT does not converge.

3.3.2 Related Research

Our approach should not be confused with the TectoMT submission by Zdeněk Žabokrtský with a deterministic transfer: heuristics fully exploiting the similarity of English and Czech t-layers.

Ding and Palmer (2005) improve over word-based MT baseline with a formalism very similar to STSG. Though not explicitly stated, they seem not to encode frontiers in the treelets and allow for adjunction (adding siblings), like Quirk et al. (2005), which significantly reduces data sparseness.

Riezler and III (2006) report an improvement in MT grammaticality on a very restricted test set: short sentences parsable by an LFG grammar without back-off rules.

4 Conclusion

We have presented our best-performing factored phrase-based English-to-Czech translation and a highly experimental complex system with tree-based transfer at a deep syntactic layer. We have discussed some of the reasons why the phrase-based MT currently performs much better.

References

Ondřej Bojar and Martin Čmejrek. 2007. Mathematical Model of Tree Transformations. Project EuroMatrix - Deliverable 3.2, ÚFAL, Charles University, Prague.

Ondřej Bojar, Zdeněk Žabokrtský, Pavel Češka, Peter Beňa, and Miroslav Janíček. 2008. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proc. of LREC 2008*. ELRA.

Ondřej Bojar. 2007. English-to-Czech Factored Machine Translation. In *Proc. of ACL Workshop on Statistical Machine Translation*, pages 232–239, Prague.

Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger . In *Proc. of ANLP-NAACL*.

Peter F. Brown et al. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Pavel Češka. 2006. Segmentace textu. Bachelor's Thesis, MFF, Charles University in Prague.

Silvie Cinková et al. 2004. Annotation of English on the tectogrammatical level. Technical Report TR-2006-35, ÚFAL/CKL, Prague, Czech Republic.

Michael Collins. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proc. of ACL*.

Yuan Ding and Martha Palmer. 2005. Machine Translation Using Probabilistic Synchronous Dependency Insertion Grammars. In *Proc. of ACL*.

Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum, Prague.

Jan Hajič et al. 2006. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4.

Liang Huang, Kevin Knight, and Aravind Joshi. 2006. Statistical Syntax-Directed Translation with Extended Domain of Locality. In *Proc. of AMTA*, Boston, MA.

Václav Klimeš. 2006. *Analytical and Tectogrammatical Analysis of a Natural Language*. Ph.D. thesis, ÚFAL, MFF UK, Prague, Czech Republic.

Jan Kocek, Marie Kopřivová, and Karel Kučera, editors. 2000. *Český národní korpus - úvod a příručka uživatele*. FF UK - ÚČNK, Praha.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proc. of EMNLP*.

Philipp Koehn, Hieu Hoang, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL Demo and Poster Sessions*.

Ryan McDonald et al. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proc. of HLT/EMNLP 2005*.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Franz Josef Och and Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In *Proc. of COLING*, pages 1086–1090.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of ACL*.

Jan Ptáček and Zdeněk Žabokrtský. 2006. Synthesis of Czech Sentences from Tectogrammatical Trees. In *Proc. of TSD*, pages 221–228.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proc. of ACL*, pages 271–279.

Stefan Riezler and John T. Maxwell III. 2006. Grammatical Machine Translation. In *Proc. of HLT/NAACL*.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia, Prague.

Zdeněk Žabokrtský. 2008. Tecto MT. Technical report, ÚFAL/CKL, Prague, Czech Republic. In prep.