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

Ondřej Bojar and Jan Hajič
Institute of Formal and Applied Linguistics
UFAL 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.

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

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

2We 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

---

143
Proceedings of the Third Workshop on Statistical Machine Translation, pages 143–146, Columbus, June 2008. ©2008 Association for Computational Linguistics
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 (Koceck 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 (Šgall 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 autonomous 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.

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:

$$\hat{\delta} = \arg\max_{\delta \text{ s.t. source is } T_1} \exp\left( \sum_{m=1}^{M} \lambda_m h_m(\delta) \right)$$

The key feature function $h_m$ in STSG represents the probability of attaching pairs of dependency treelets $t_{1:2}^i$ such as in Figure 1 into aligned pairs of frontiers ($\circ$) in another treelet pair $t_{1:2}^j$ given frontier state labels (e.g. $\preceq_{\text{Pred}}-\text{VP}$ in Figure 1):

$$h_{STSG}(\delta) = \log \prod_{i=0}^{k} p(t_{1:2}^i | \text{frontier states})$$

Other features include e.g. number of internal nodes (drawn as • in Figure 1) produced, number of treelets produced, and more importantly the traditional $n$-gram language model if the target (a-)tree
is linearized right away or a binode model promoting likely combinations of the governor \( g(e) \) and the child \( c(e) \) of an edge \( e \in T_2 \):

\[
h_{\text{binode}}(\delta) = \log \prod_{e \in T_2} p(c(e) | g(e)) \tag{3}\]

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

\(^3\)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.

<table>
<thead>
<tr>
<th>Tree-based Transfer</th>
<th>LM Type</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>epcp</td>
<td>n-gram</td>
<td>10.9±0.6</td>
</tr>
<tr>
<td>eaca</td>
<td>n-gram</td>
<td>8.8±0.6</td>
</tr>
<tr>
<td>epcp</td>
<td>none</td>
<td>8.7±0.6</td>
</tr>
<tr>
<td>eaca</td>
<td>none</td>
<td>6.6±0.5</td>
</tr>
<tr>
<td>etca</td>
<td>n-gram</td>
<td>6.3±0.6</td>
</tr>
<tr>
<td>etct, factor preserving structure</td>
<td>binode</td>
<td>5.6±0.5</td>
</tr>
<tr>
<td>etct, factor preserving structure</td>
<td>none</td>
<td>5.3±0.5</td>
</tr>
<tr>
<td>eact, target side atomic</td>
<td>binode</td>
<td>3.0±0.3</td>
</tr>
<tr>
<td>etc, atomic, all attributes</td>
<td>binode</td>
<td>2.6±0.3</td>
</tr>
<tr>
<td>etc, atomic, all attributes</td>
<td>none</td>
<td>1.6±0.3</td>
</tr>
<tr>
<td>etc, atomic, just t-lemmas</td>
<td>none</td>
<td>0.7±0.2</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>WMT07</th>
<th>WMT08</th>
</tr>
</thead>
<tbody>
<tr>
<td>DevTest</td>
<td>NC Test</td>
</tr>
<tr>
<td>Moses</td>
<td>14.9±0.9</td>
</tr>
<tr>
<td>Moses, CzEng data only</td>
<td>13.9±0.9</td>
</tr>
<tr>
<td>etct, TectoMT annotation</td>
<td>4.7±0.5</td>
</tr>
</tbody>
</table>

Table 2: WMT08 shared task BLEU scores.

rules for t-layer parsing and generation instead of Klimeš (2006) and (Ptáček and Žabokrtský, 2006).

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.
Lack of \emph{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 $\lambda_m$ 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.


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


