Towards English-to-Czech MT via Tectogrammatical Layer

O. Bojar, S. Cinková, and J. Ptáček
Outline

• Overview
• Synchronous Tree Substitution Grammar
• Experimental Results
• Discussion
Overview: Deep Syntatic Machine Translation

Diagram:
- Interlingua
  - Tectogrammatical (t-) Layer
  - Analytical (a-) Layer
  - Morphological (m-) Layer
- English
- Czech
- etct
- generate
Under the Hood

- Parallel treebank
- Little Trees Pairs Dictionary
- Synchronous Tree Substitution Grammar

- Parsing
- Covering the source tree with treelets with back-offs
- Reading the target tree
- Generation
Synchronous Tree Substitution Grammar

• (Hajič, 2002)
• (Eisner, 2003)
• (Čmejrek, 2006)
• Not bounded to t-trees
Idea: Observe a Pair of Dependency Trees

# Asociace uvedla, že domácí poptávka v září stoupla.

# The association said domestic demand grew in September.
Idea: Decompose Trees into Treelets

# Asociace uvedla, že domácí poptávka v září stoupla.

# The association said domestic demand grew in September.
Idea: Collect Dictionary of Treelet Pairs

_Sb_{cs} uvedla, že _Pred_{cs} = _Sb_{en} said _Pred_{en}

asociace = The association

_Adj_{cs} poptávka = _Adj_{en} demand

domestic = domácí
Little Trees Formally

Given a set of states $Q$ and a set of word labels $L$, we define:

A little tree or treelet $t$ is a tuple $(V, V^i, E, q, l, s)$ where:

- $V$ is a set of nodes,
- $V^i \subseteq V$ is a nonempty set of internal nodes. The complement $V^f = V \setminus V^i$ is called the set of frontier nodes,
- $E \subseteq V^i \times V$ is a set of directed edges starting from internal nodes only and forming a directed acyclic graph,
- $q \in Q$ is the root state,
- $l : V^i \rightarrow L$ is a function assigning labels to internal nodes,
- $s : V^f \rightarrow Q$ is a function assigning states to frontier nodes.

Optionally, we can keep track of local or global ordering of nodes in treelets.

I depart from Čmejrek (2006) in a few details, most notably I require at least one internal node in each little tree.
Treelet Alignments: Heuristics

- Similar to common phrase-extraction techniques given word alignments.
- Basic units are little trees instead of word spans.

1. Obtain node-to-node alignments (GIZA++ on linearized trees).

2. Extract all treelet pairs satisfying these conditions:
   - no more than $i$ internal nodes and $f$ frontier nodes,
   - compatible with node alignment,
     e.g. no node-alignment link leads outside the treelet pair and frontiers are linked.
   - satisfying STSG property.
     All children of an internal node have to be included in the treelet (as frontiers or internals),
     i.e. assume no adjunction operation was necessary to construct the full tree.

3. Estimate probabilities, e.g. $p(t_1, t_2 | \text{root state}_1, \text{root state}_2)$
Back-off Schemes

Preserve all. Full-featured treelets are collected in training phase.
Required treelets often never seen in training data ⇒ back-off needed.

Drop frontiers. Observed treelets reduced to internal nodes only.
Given a source treelet, internals translated by the dictionary, frontiers generated
on the fly, labelled and positioned probabilistically.

Word for word. Useful for single-internal treelets only: The label of the root
internal translated independently, frontiers generated on the fly, labelled
probabilistically, order unchanged.

Keep a word non-translated to handle unknown words.
Allowed only for single-internal treelets, frontiers mapped probabilistically.
Decoding STSG

- Find target tree such that the synchronous derivation $\delta$ is most likely.
- Implemented as two-step top-down beam-search similar to Moses:

1. Prepare **translation options table**:
   - For every source node consider every subtree rooted at that node.
   - If the subtree matches the source treelet in a treelet pair, we've got a translation option.
   - Keep only best $\tau$ translation options at a node.

2. Gradually **expand partial hypotheses**:
   - Starting at root use translation options to cover source tree.
   - Keep only best $\sigma$ partial hypotheses of a given size (input nodes covered).
Implementation Details

- STSG model extended to log-linear combination of features:

$$\hat{\delta} = \arg\max_{\delta \in \Delta(T_1)} \exp \left( \sum_{m=1}^{M} \lambda_m h_m(\delta) \right)$$

(1)

instead of

$$\hat{\delta} = \arg\max_{\delta \in \Delta(T_1)} p(t_{1:2}^0 | \text{Start}_{1:2}) \times \prod_{i=1}^{k} p(t_{1:2}^k | q_{1:2}^k)$$

(2)

- Tincodb (like GDBM) to store and access treelet dictionaries.
- Target tree structure can be disregarded (output linearized right away).
  - IrstLM to promote hypotheses containing frequent trigrams.
- Implemented in Mercury (Somogyi, Henderson, and Conway, 1995).
- Parallel computation on Sun Grid Engine cluster (160 CPUs in 40 machines).
Generator

• Deterministic
• First plan - then fulfil
• Expects full featured t-trees on input
Overview: Deep Syntactic Machine Translation

Tectogrammatical (t-) Layer
Analytical (a-) Layer
Morphological (m-) Layer

Interlingua

Generate
Linearize

---

English
Czech
## Empirical Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>eact</td>
<td>3.0 ± 0.3</td>
</tr>
<tr>
<td>etcct</td>
<td>5.0 ± 0.5</td>
</tr>
<tr>
<td>etca</td>
<td>6.3 ± 0.6</td>
</tr>
<tr>
<td>eaca</td>
<td>8.6 ± 0.6</td>
</tr>
<tr>
<td>epcp</td>
<td>10.3 ± 0.7</td>
</tr>
<tr>
<td>epcp with no state labels</td>
<td>11.0 ± 0.7</td>
</tr>
<tr>
<td>Phrase-based (Moses)</td>
<td></td>
</tr>
<tr>
<td>Vanilla</td>
<td>12.9 ± 0.6</td>
</tr>
<tr>
<td>Factored</td>
<td>14.2 ± 0.7</td>
</tr>
</tbody>
</table>

ACL 2007 WMT shared task data, 55k training sentences, 964 test sentences.
Upper Bound on MT Quality via t-layer

- Analyse Czech sentences to t-layer.
- Optionally ignore some node attributes.
- Generate Czech surface.
- Evaluate BLEU against input Czech sentences.

\[
\text{BLEU} \\
\begin{align*}
\text{Full automatic t-layer, no attributes ignored} & \quad 36.6 \pm 1.2 \\
\text{Ignore sentence mood (assume indicative)} & \quad 36.6 \pm 1.2 \\
\text{Ignore verbal fine-grained info (resultativeness, \ldots)} & \quad 36.6 \pm 1.2 \\
\text{Ignore verbal tense, aspect, \ldots} & \quad 24.9 \pm 1.1 \\
\text{Ignore all grammatical} & \quad 5.3 \pm 0.5
\end{align*}
\]

⇒ Node attributes obviously very important.