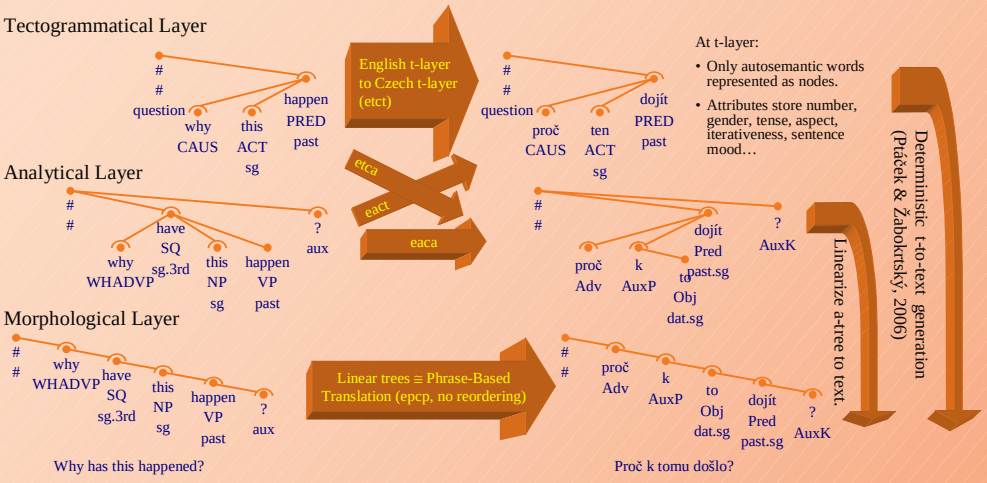
Phrase-Based and Deep Syntactic English-to-Czech SMT

Ondřej Bojar, Jan Hajič, {bojar,hajic}@ufal.mff.cuni.cz, Charles University in Prague, MFF, ÚFAL

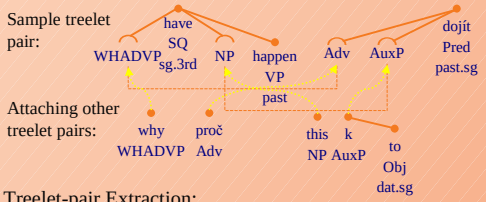
Transfer at Various Layers



Synchronous Tree Derivation

Synchronous Tree-Substitution Grammars (STSG, Eisner, 2003)

1. Decompose source tree into treelets.
2. Translate treelets.
3. Join target-side treelets attaching treelet roots to frontier nodes.
4. Linearize a-tree or generate plaintext from t-tree.



Treelet-pair Extraction:

1. Annotate sentence-parallel corpus up to t-layer:
   - **Automatic:** TectoMT environment (Žabokrtský, 2008) utilizes various taggers, parsers…
   - **Manual:** Prague Czech-English Dependency Treebank (PCEDT 2.0, in progress)
2. GIZA++ to obtain node-to-node alignments.
3. Extract all treelet pairs compatible with node alignment.

Search for the most likely synchronous derivation:

Stack-based top-down beam search similar to Moses.
Two stages:

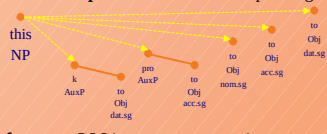
1. Generate translation options (target-side treelets). Various back-off methods, e.g. frontiers disregarded and generated on the fly. Output attributes generated as in factored translation (mapping and generation steps).
2. Attach translation options to frontiers.

Log-linear model of the following features:

- STSG model: *p(treelet pair | frontier states)*
- Direct and reverse translation models: $\Pi_{\text{treelet pairs}}\ p(\text{target treelet} \mid \text{source treelet})$
- Direct and reverse models for factored attribute generation.
- n-gram language model for e*ca and e*cp (trees directly linearized)
- Binode language model for e*ct: $\Pi_{\text{edge } e}\ p(\text{child}(e) \mid \text{governor}(e))$
- Number of internal nodes covered…

Problems of Deep Transfer

Cumulation of errors at every step of analysis (2×tagging, 2×parsing to a-layer, rules/parsing to t-layer).
Data loss at treelet pair extraction: natural divergence and annotation errors ⇒ incompatible tree structures and node alignment ⇒ many treelet pairs not extracted.
Combinatorial explosion in translation options generation:



There are around 20 node attributes in t-trees.
Treated as atomic, t-node labels have higher entropy (11.54) than lowercase plaintext (10.74). The idea is that the attributes should be more or less independent and should map easier across languages.
Current implementation requires *fully generated* target treelets
⇒ too many options for target node attributes given little context
⇒ too many similar target treelets created
⇒ e.g. different lexical choices pushed off the stack
⇒ hypotheses on n-best list differ in less relevant attributes only.

Lack of n-gram LM in t-to-text generation.
We support final LM-based rescoring but there is too little variance in n-best lists.
Too many model parameters, esp. with factored output nodes ⇒ MERT fails to coverge.

Empirical Results

BLEU scores for various tree-based transfer configurations:
WMT 07 DevTest, automatic t-layer by Klimeš (2006), compared to phrase-based MT (Bojar, 2007).

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

BLEU scores for our WMT 08 submissions:
WMT 07 DevTest and WMT 08 tests, compared to etct (factored, preserving structure) with automatic t-layer by Žabokrtský (2008).

	WMT 07	WMT 08	
	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

Phrase-Based Setup

Moses configuration for English-to-Czech translation:

Parallel corpus: CzEng 0.7, about 1M parallel sentences.
Word alignment: GIZA++ on Czech and English lemmas.
Truecasing: Uppercased names preserved, sentence capitalization removed.
Decoding steps:
English truecased form → Czech truecased form + 3×LM
Czech morphological tag + 3×LM
Language models:

- 3-grams of word forms (CzEng target side, 15M tokens),
- 3-grams of word forms (NC Test domain, 1.8M tokens),
- 4-grams of word forms (Czech National Corpus, 365M tokens),
- 3×7-gram models of morphological tags (same data sources).

Lexicalized reordering using monotone/swap/discontinuous bidirectional model on source and target word forms.
Minimum error-rate training (MERT) optimizing for BLEU.

Summary

- Implemented a complex syntax-based system for English-to-Czech MT.
- STSG top-down decoder applied at various layers of language description.
- Significant improvement of "etct" using various methods of back-off, including factored translation of node attributes.

However:

- The more complex setup, the worse BLEU scores due to cumulation of errors, data loss and combinatorial explosion (⇒search errors).
- Best English-to-Czech quality currently achieved by factored phrase-based MT with a big target-side LM.

Full text, acknowledgement and the list of references in the proceedings of ACL WMT 08.