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

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

Transfer at Various Layers

Tectogrammatical Layer

Analytical Layer

Morphological Layer

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.

Sample treetree pair:

Treelet-pair Extraction:
1. Annotate sentence-parallel corpus up to t-layer:
   - Automatic: TectoMT environment (Zabokrtský, 2008)
   - 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.

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 treetreelet pair extraction: natural divergence and annotation errors ⇒ incompatible tree structures and node alignment ⇒ many treetreelet pairs not extracted.

Combinatorial explosion in translation options generation:

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 converge.

Empirical Results

BLEU scores for various tree-based transfer configurations:

| WMT07 Dev/Test and WMT08 tests, compared to phrase-based MT (Bojar, 2007) |
|-----------------|-----------------|-----------------|
|                 | Phrase-Based     | Deep Syntactic |
|                 | Transfer         | Transfer        |
|                 | LM Type          | BLEU            | BLEU            |
| WMT07 Dev/Test  | n-gram           | 10.9±0.6        | 10.9±0.6        |
| WMT08 Test      | n-gram           | 8.8±0.6         | 9.6±0.6         |
| Moses           | Moses            | 6.3±0.6         | 6.3±0.6         |
| CzEng data      | Moses            | 0.7±0.3         | 0.7±0.3         |

Phrase-Based Setup

Moses configurations for English-to-Czech translation:

Parallel corpus: CzEng07, about 1M parallel sentences.

Word alignment: GIZA++ on Czech and English lemmas.

Truecasing: Uppercased names preserved, sentence capitalization removed.

Decoding steps:
1. English truecased form -> Czech truecased form
2. Phrase-Based Setup

Language models:
1. 3-grams of word forms (CzEng target side, 15M tokens).
2. 3-grams of word forms (NC Test domain, 1.8M tokens).
3. 4-grams of word forms (Czech National Corpus, 365M tokens).
4. 3-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 “ect” using various methods of back-off, including factored translation of node attributes.
- The more complex setup, the worse BLEU scores due to cumulative errors, data loss and combinatorial explosion of 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.