New Language Pairs in TectoMT

Ondřej Dušek, Luís Gomes, Michal Novák, Martin Popel, and Rudolf Rosa

{odusek,mnovak,popel,rosa}@ufal.mff.cuni.cz
Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
luis.gomes@di.fc.ul.pt
University of Lisbon, Faculty of Sciences, Department of Informatics

Introduction

• TectoMT – competing in WMT since 2008, just English-Czech
• New language pairs added in the QTLeap project:
  • Dutch, Spanish, Basque, Portuguese (to and from English)
  • Czech-English – new language pair in WMT'15

New Languages: Standards & Training

• TreeX blocks for new languages made easier:
  • Common morphology – Interset
  • Common syntactic style – HamleDT 1.5 (3.0 / UD planned)
  • Base language-independent blocks
  • Makefiles for easy translation model training

TectoMT System Operation

• Analysis – transfer – synthesis
  • a-layer – dependency trees
    • one node per token
  • t-layer – deep syntactic trees
    • only content words have nodes
    • t-lemma: deep lemma
    • formeme: morpho-syntactic function label
    • grammatemes: grammatical meaning

Transfer in TectoMT

• Translating t-layer trees node-by-node (assuming same shape)
• Factorized (different models):
  • t-lemma, formeme – discriminative (MaxEnt) models
  • grammatemes – rules
• Several options provided by each model
• Hidden Markov Tree Model – selecting the best combination of model outputs

New in WMT’15: Czech-English Translation

Analysis

• Modified from CaEng, English-Czech TectoMT training
• MorphoDiTa tagger
• MST parser adapted for Czech

Synthesis

• Gradually transform a copy of the translated t-tree

1. Fill in morphological attributes
(based on grammatemes)
2. Mark subject (for agreement)
3. Enforce basic word order
4. Enforce subject-predicate agreement
5. Add auxiliary words from formeme
(prepositions, conjunctions)
6. Add articles
7. Add auxiliary verbs
8. Remove imperative subjects
9. Add negation particles
10. Add punctuation
11. Inflex words (Morphodita + Flect)
12. Transform a -> an
13. Delete repeated prepositions and conjunctions in coordinations
14. Capitalize first word in sentence

WMT’15 Results

• English-Czech: 13.9% BLEU – among the last, human eval. same
• used in Chimera, overall winner in automatic+human ranking
• 8.0% of reference tokens are only in TectoMT (and not Moses)
• more than half of these tokens were used in Chimera
• TectoMT is essential for Chimera’s success
• Czech-English: 12.8% BLEU – last, human eval. 2nd-to-last
• pruning was too easy (bug)

New Languages: Simple Model Combination

• Combining t-lemma + formeme predictions
• No training required, as opposed to Hidden Markov Tree Model
• Five non-parametric functions combining the models’ outputs:
  • AM-P – arithmetic mean of model-predicted probabilities
  • GM-P – harmonic mean of probs
  • GM-Log-P – geometric mean of log-probs
  • HM-Log-P – harmonic mean of log-probs
• Tested on QTLeap corpus
• For 3 language pairs
• Performance better than baseline
  (using 1st variant of lemma and formeme)
• HMTM is better, if available