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

- Tree 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
  - forname: semantic/syntactic function label
  - grammemes: morpho-syntactic function label
  - grammameses: grammatical meaning

Transfer in TectoMT

- Translating t-layer trees node-by-node (assuming same shape)
- Factorized (different models):
  - t-lemma, forname – discriminative (MaxEnt) models
  - grammameses – 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

Transfer

- Basic t-lemma + forname models with rule-based overrides
- Czech gender removed
- Double negatives removed
- Name translation fixes
- Grammatamese fixes, e.g.: těstoviny (pl) -> pasto (sg)

Synthesis

- Gradually transform a copy of the translated t-tree
  1. Fill in morphological attributes (based on grammameses)
  2. Mark subject (for agreement)
  3. Enforce basic word order
  4. Enforce subject-predicate agreement
  5. Add auxiliary words from forname (prepositions, conjunctions)
  6. Add articles
  7. Add auxiliary verbs
  8. Remove imperative subjects
  9. Add negation particles
  10. Add punctuation
  11. Inflect words (Morphodita + Flect)
  12. Transform o -> 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 80% 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 eager (bug)

Conclusion and Future Work

- Improvements and bugfixes required
  - Hidden Markov Tree Model for Czech
  - word order fixes, article assignment (English)
- Further development plans:
  - Interset instead of grammameses on t-layer, Universal Dependencies
  - Vowpal Wabbit and word embedding features in transfer models
- possibilities of non-isomorphic transfer

Presented at WMT 2015, Lisbon, September. Supported by the FP7 EU grant QTLeap (No. 610516), and SVV project 260 104 and GAUK grants 2582014 and 338915 of the Charles University in Prague.
Using language resources hosted by the LINDAT/CLARIN Research Infrastructure, Project No. LHG180110 of the Czech Ministry of Education, Youth and Sports.