DEPFIX: A System for Automatic Correction of Czech MT Outputs

Rudolf Rosa, David Mareček, Ondřej Dušek; \{rosa, marecek, odusek\}@ufal.mff.cuni.cz

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL)

Goal: Improve Czech grammar

- Phrase-based MT often breaks morphological agreement
- Some of the dependencies still recovered in parses of MT output
  ⇒ We can check and fix the agreements.

DEPFIX

- Rule-based grammar correction of English-to-Czech MT
- Applicable on top of any MT system
- WMT12: cu-depfix = onlineB + Depfix
  Input: source (English sentences), target (Czech MT output)
  1. Tag source and target (morphological tags and lemmas)
  2. Align source with target using lemmas (GIZA++)
  3. Parse source (MST Parser)
  4. Parse target (RUR Parser)
  5. Apply Depfix rules
    Output: corrected target

RUR Parser

- Non-projective two-stage dependency parser
- A modification of the MST parser (McDonald, 2005)
- Adapted for parsing of SMT outputs:
  • Features projected from source sentence analysis
  (tag, dependency relation, dependency label)
  • “Worsening” the training treebank
    (changing correct fiction of some words to incorrect based on
    a tag error model estimated on Moses output)
  • Labeling done in a second stage (MIRA-based labeler, adapted
    in a similar way as the parser)

Rules

- Preposition “by”
  • “by + noun” is translated by putting the noun into:
    - the instrumental case, if modifying a verb
    - the genitive case, if modifying a noun
- Subject case
  - Subject must be in the nominative case
- Subject-Past Participle agreement
  - Subject and past participle must agree in number and gender
- Passive–Auxiliary “be” agreement
  • Aux. verb “to be” must agree with passive in gender and number
- Preposition–Noun agreement
  - Nouns must agree with prepositions in morphological case
- Preposition “of”
  • “of + noun” is translated by putting the noun into the genitive case
- Noun-Adjective agreement
  - Adjectives must agree with nouns in number, gender, and case