Generation in MT

Jan Hajič
The Team

– Senior members & affiliate members
  • Jan Hajič, Charles Univ., Prague
  • Drago Radev, Univ. of Michigan
  • Gerald Penn, Univ. of Toronto
  • Jason Eisner, Johns Hopkins Univ.
  • Owen Rambow, Univ. of Pennsylvania
  • Dan Gildea, Univ. of Pennsylvania
  • Bonnie Dorr, Univ. of Maryland

– Students:
  • Yuan Ding, Univ. of Pennsylvania
  • Martin Čmejrek, Charles Univ., Prague
  • Terry Koo, MIT
  • Kristen Parton, Stanford Univ.
The Goal

- Generate English (plain surface form)
  - from syntactic-semantic sentence representation (so-called “tectogrammatical”, or TR)
- Possible application setting:
  - machine translation
  - other uses:
    - part of front-end for QA systems, full generation
- Evaluate under various circumstances
The Framework

- “Classic” MT design assumed
  - Analysis - Transfer - Synthesis
- Tectogrammatical level at transfer stage
  - Dependency syntactic-semantic representation
- Language pair:
  - from Czech to English
The Framework

source language text

morphology/tagging

surface syntax

deep syntax (tectogrammatics)

transfer

depth syntax to surface syntax, word order

lemma+tag generation

morphology (gen.)

target language text

transfer

7/15/2002

WS02: Generation in MT: Intro Day
According to his opinion UAL's executives were misinformed about the financing of the original transaction.
For comparison: Surface dependency

3/51: #3 According to his opinion UAL’s executives were misinformed about the financing of the original transaction.
According to his opinion UAL’s executives were misinformed about the financing of the original transaction.
WS02 Task

• Generation [in an MT framework]
  – from: deep tectogrammatical representation
  – to: surface plain text form
  • via:
    – surface syntax representation (dependency/parse tree)
    – lemmas+tags (plus morphological synthesis)

• Key issues
  – lexical choice; “auxiliaries”; word order; morphology
WS02 Data + Tools Available

- Translation of the Penn Treebank to Czech (250k)
- Czech analysis (text → tectogrammatical representation)
  - tagger
  - parser (surface syntax, from WS’98)
  - deep parser (rule-based)
- Translation dictionary (multiple translations, non-prob.)
- Tectogrammatic annotation of English (Penn TB data)
- English morphology (morpha: adapted, cleaned)
- Evaluation script
Generation: Complex Input

English translation:
act issue proceeds return
Evaluation

• Simple and measurable
  • BLEU (IBM): 5 variants, 250+250 sentences

• Two (three?) “tracks”:
  • baseline from TR English (manually annotated)
  • complete translation from Czech
  • ?from Proposition Bank (comparison to TR)

• Generation: rule based vs. statistical

• Does it work?
  • Compare also to word-based statistical MT (WS’99)