Generation in MT

Progress Report #1

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 Motivation

• Tectogrammatical Representation
  – linguistic intuition:
    • TR best represents structure & (linguistic) meaning
    – “best”: using compact description, as abstract as possible (wrt surface syntax, phrase structure)
    – => can be shared to a large extent among languages
    – => needs less data to train statistical models
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                             target language text

morphology/tagging  transfer

deep syntax (tectogrammatics)  deep syntax to surface syntax, word order

surface syntax  lemma+tag generation

morphology (gen.)

source language text  target language text

WS’02
According to his opinion UAL’s executives were misinformed about the financing of the original transaction.
Progress so far (1)

• We have
  – Data and Czech-side tools at CLSP
    • tools working up to transfer (but without it yet)
  – looked at Cze/Eng TR/AR trees
  – spent 120 man-hours discussing how to do it
    • form of the model: source channel vs. classifier
    • features to be used
  – got data & info for conversion of Penn PropBank
Progress so far (2)

• We have
  – compared the Eng TR trees with what’s needed for the symbolic method’s input
  – compared Eng TR trees: auto vs. manual
  – experimented with the Eng word-order model
  – working code for Eng morphology
  – working code and scripts for the evaluation experiments proper
Division of Labor: Pipeline

- Local tree substitution (twisting & translation) via dynamic prog.
- Decisions requiring global information via classifiers
- Any remaining node ordering decisions via language model

Optional diagram:

- Must be parameterized use undirected GM?
Model Architecture

Noisy channel

Source model

Classifiers

Another source model???
Source model for first-stage approximation

Deterministic channel

Result constrains source model

Current Compromise

Noisy channel

Source model for second stage (only deals with ordering)
English TR trees: auto vs. manual

• Wrote specific “diff” tools (YD)
• Test: 284 sentences, sect. 17 (5081 nodes)
  – TR lemma match: 93.4%
  – Functor match: 79.4%
  – Dependency (structural) match: 88.3%
  – Swapped dependencies: 36 (< 1%)
**PropBank Input**

- **PropBank**: Penn project, predicate-argument
- **Goal**: use PropBank to
  - Improve automatic construction of English TRs
  - Allow generation from “generic” pred-arg structures
- **Tasks**
  - Augment PropBank with roleset info $\sqrt{\Box}$
  - Add lexical-conceptual role tags
  - Convert to TG (following Hajičová & Kučerová)
Word order

• Three major experiments (DG):
  – Tree-based models:
    • Collins model on PennTB style (parse) trees
      – 97% words at correct position reconstructed
    • Analytical level surface dependency, tree-based
      – 94% (chance: 68%)
        • levels >= 7 nodes ignored: 1.5% of nodes abs.
  – Bigram surface model, PennTB style trees
    – dynamic programming (begin/end words of phrase)
      – 86% (chance: 64%)
English Morphology

• Data (JC)
  – WSJ (35 MW) analyzed by the “morpha” tool
    – PennTB compatible tagset + lemmatization
  – table extracted (659843 entries), some corrected:
    
    |  |  |
    |---|---|
    | VBG | opening | open | 14049 |
    | NNS | openings | opening | 931 |
    | NN  | opening  | opening | 12084 |

• Code (KP)
  – table lookup (KP), accepts multiple formats
  – currently correcting dictionary, overall error rate computation
Morphology Coverage
Symbolic approaches

• FUF/Surge (Elhadad/Robin): almost everything needs to be specified
• Nitrogen/Halogen (Langkilde-Geary/Knight): less specification is OK, uses statistical reranker
Pierre Vinken, 61 years old, will step down as nonexecutive director of the board Nov. 26.
Evaluation

• Evaluation scripts (TK)
  – Multiple dimensions of evaluation, presentation
  – Core software: BLEU
    • by and from Kishore Papineni
  – Format conversion
    • from morphology output
    • from translations
  – Status: three reference translations so far