Consistency and Parsing in the Arabic Treebank

Seth Kulick
skulick@seas.upenn.edu
Overview

◆ Two topics: Consistency and Parsing
  ● Effect of treebank revisions on parsing
  ● Verifying consistency of a treebank directly (i.e., without using parser)

◆ Quality Control and InterAnnotatorAgreement
  ● What we’ve done so far

◆ Impact of completed revisions on parsing

◆ Quality Control and InterAnnotatorAgreement
  ● Limits of approach, current/future work
Current QC and IAA

- Local tree-fragment checking
  - Greater use of Corpus Search
  - Some additional token/pos changes based on tree
- (Limited) global word-based changes
  - Tokenization/pos revisions of “function words”
    - Two levels of annotation – source and tree tokens
    - Internal reorganization of treebank
- InterAnnotator Agreement
  - Focus on number(s) from evalb (94.3)
Parsing Experiment

◆ New ATB and old ATB:
  • Parsed ATB1,2,3 separately and ATB123 together
  • Mona Diab’s train/dev/test split (<=40 words)
  • Using gold tokenization/tags - A serious limitation
  • Two modes
    ◆ Parser uses its own tags for “known” words
    ◆ Parser forced to use given tags for all words
  • LDC reduced TAG set (+DET)

◆ Penn Treebank
  • Made up training, test same size as ATB3, 123
- Nice improvement, not at PTB level yet
- Results not as good for test section
- Dependency Analysis shows:
  - Improvement in recovery of core syntactic relations
  - Problem with PP attachment!

(Kulick, Gabbard, Marcus TILT 2006, Gabbard & Kulick 2008 ACL)
Parsing Analysis
Core Syntactic Relations Recovery

- Across the board improvement

- Attachment issues remain

---

**Bar Chart**

- ATB123-old
- ATB123-new
- PTB

- NP->noun NP
- PP->prep NP
- baseNP
- VP->verb NP
- SBAR->compl S
- S->NP VP

---

**Bar Chart**

- ATB123-old
- ATB123-new
- PTB

- VP->verb PP
- NP->baseNP PP
- NP->NP PP

---

Ontonotes – 2/24/09 6/9
In July 2005 Turkey had signed A protocol.
In July 2005, Turkey had signed a protocol.
(Near) Future QC and IAA – Methodology

Treebank Decomposition

- Long history for NLP, massively underutilized for corpus creation (Xia, 2001)

In July 2005 Turkey signed a protocol
(Near) Future QC and IAA - Benefits

- Correlation with all different levels of treebank
  - From source tokens downstream to tree contexts
  - Allows new types of searches
- More meaningful IAA analysis
  - Decompose annotators’ trees
  - Useful automatic information:
    - e.g. What core structures are they agreeing on?
- Also fun for parsing, parsing analysis
  - Decompose gold tree and parse output, compare
- Even for MT eval (Kulick & Marcus)
Concluding Thoughts

◆ Consistency and Parsing
  ● Parser results suggest improved consistency
  ● More to do to guarantee consistency w/o parsing

◆ What other kinds of annotation do we need?
  ● +/- human : MASC PL nonhuman == FEM SG
  ● Verb pattern information

◆ More feedback please
  ● Take advantage of analysis for release format? heads, better source/tree token representation?