Treebanks

Judith L. Klavans

Three Points

1. The value of “Agree to Disagree”
2. Different folks, different strokes
   - How to use treebanks in Evaluation if what you’re evaluating is different from the established standard?
   - How to map across representations?

1. Plus ça change – The Old and the New

The value of Agree to Disagree

E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski.

A procedure for quantitatively comparing the syntactic coverage of English grammars. 1991. Procedings of DARPA Speech and Language Workshop.

- “The problem of quantitatively comparing the performance of different broad-coverage grammars of English has to date resisted solution.
- *Prima facie*, known English grammars appear to disagree strongly with each other as to the elements of even the simplest sentences.
- For instance, the grammars of Steve Abney (Bellcore), Ezra Black (IBM), Dan Flickinger (Ilewlett Packard), Claudia Gdaniec (Logos), Ralph Grishman and Tomek Strzalkowski (NYU), Phil Harrison (Boeing), Don Findle (AT&T), Bob Ingria (BBN), and Mitch Marcus (U. of Pennsylvania) recognize in common only the following single constituents, when each grammarian provides the single parse which he/she would ideally want his/her grammar to specify for three sample Brown Corpus sentences….”

“Big Issues”

- treatment of punctuation as independent tokens or, on the other hand, as parasites on the words to which they attach in writing;
- the recursive attachment of auxiliary elements to the right of Verb Phrase nodes, versus their incorporation there *en bloc*;
- the grouping of pre-infinitiva1 "to" either with the main verb alone or with the entire Verb Phrase that it introduces; and
- the employment or non-employment of "null nodes" as a device in the grammar;
- as well as other differences.

Despite the seeming intractability of this problem, it appears to us that a solution to it is now at hand.

We propose an evaluation procedure with these characteristics:

- judge a parse based only on the constituent boundaries it stipulates (and not the names it assigns to these constituents)
- compare the parse to a "hand-parse" of the same sentence from the University of Pennsylvania Treebank
- two principal measures for each parse submitted
  - Precision
  - Recall
- Who knew?

2. Different folks, different strokes

- Theory-neutral treebanks:
  - do not adhere to any particular linguistic theory
  - encode those grammatical properties that are distinguished by many, if not all grammatical frameworks
- Advantage:
  - More widely usable
  - Less dependent on whatever version of a particular grammatical theory may have existed at the time when the treebank annotation scheme was determined
- Examples: Penn Treebank, Negra treebank, Tübingen treebanks

Theory-Neutral and Theory-Supporting TB

- attempt to combine the advantages of theory-neutral and theory-specific treebank annotation.
- target annotation schemes are theory-specific
- the source annotation scheme must at least be neutral in the sense that it supports conversion to all target schemes.

Advantages of “Theory-Neutral”

- should allow us to produce a number of theory-specific treebanks at substantially lower cost than if each treebank had to be developed independently
- should allow us to make systematic comparisons between analyses couched in different theoretical frameworks.
- connections with work on grammar conversion for evaluation purposes (Kinyon and Rambow 2003)

Theory-neutral vs. Theory-dependent?

➔ Every decision is a theoretical decision

- No such thing as “theory-neutral”
- If you are interested in particular theory (or not), these treebanks are extremely useful
- Encoding a grammar into a treebank provides excellent feedback on the theory

A few theory-specific Treebanks

- Prague Dependency Treebank
  - based on Dependency Grammar
- The Redwoods HPSG Treebank
  - based on Head-Driven Phrase Structure Grammar
- CCGbank
  - translation of the Penn Treebank into a corpus of Combinatory Categorial Grammar derivations

- Arabic - Penn Arabic Treebank
- Bulgarian
  - HPSG-based Syntactic Treebank of Bulgarian (BulTreeBank)
- Catalan
  - CAT3LB project
- Czech
  - Prague Dependency Treebank

- Verbmobil Treebank of Spoken Japanese (Tu¨Ba-J/S)
- Portuguese
  - The Floresta Sinta(c)tica project
- Swedish
  - Talbanken05, Swedish Treebank
- Turkish
  - METU treebank

Persistent Treebank Issues

- Complete analysis vs. partial analysis
  - Syntactic *chunks* are easier to annotate more reliably
  - can be used for a variety of purposes
  - chunks are generally non-recursive NPs and PPs
- Constituency vs. dependency annotation
  - Within constituency annotation: should we annotate grammatical functions?

How to use treebanks in Evaluation if what you’re evaluating is different from the established standard?

3. Plus Ça change - The Old and the New

- Challenges
  - Constituent structure
  - Dependencies and crossing dependencies
    - Some labels more semantic, some syntactic
    - Prague Dependency – morphemic layer
  - Coordination structures
  - Discontinuities – e.g. extraposition
  - VP adjuncts - sentential & VP adjuncts

Evaluation and Computational Complexity (D. Klein et al. many papers)

- Treebank grammars can be enormous
- Raw FSA grammar may ~10K states, excluding lexicon
- Better parsers usually make the grammars larger, not smaller
- Parsing with the vanilla treebank grammar: ~ 20K Rules
- Observed exponent: 3.6
- BUT - worse in practice
- Longer sentences “unlock” more of the grammar
- All kinds of systems issues don’t scale
- Independence assumptions are often too strong.

Multilingual Corpora

- Corpus-based induction of syntactic structure: Models of dependency and Constituency, *Dan Klein and Chris Manning*, (2004) and more later.

- Goal: improve state-of-the-art monolingual natural language processing models using unannotated bilingual text
- Method: agreement between monolingual and bilingual models.

Two methods:

- monolingual view - supervised predictors learned separately for each language.
- bilingual view - log-linear predictors learned over both languages on bilingual text.
- training estimates the parameters of the bilingual model using the output of the monolingual model
- combine the two models to account for dependence between views.
- Task: named entity recognition

Results

- Bilingual predictors increases F1 by 16.1% absolute over supervised monolingual model
- Retraining on bilingual predictions increases *monolingual* model F1 by 14.6%
- For syntactic parsing, bilingual predictor increases F1 by 2.1% absolute
- Retraining a monolingual model on output - improvement of 2.0%

Future Challenges

- Mapping
  - Representations
  - Languages
  - Genres
- Multilingual Corpora
  - Mapping
  - Induction of correspondences
  - Use in Systems and Applications
- Computational Complexity

Three Points

1. The value of “Agree to Disagree”
2. Different folks, different strokes
   - How to use treebanks in Evaluation if what you’re evaluating is different from the established standard?
   - How to map across representations?
3. Plus Ça change – The Old and the New
   - Enduring Challenges
   - Complexity
   - Multilingual mapped corpora

Thank you