Conversions for heterogeneous treebank parsing (2)

L715: Seminar on: Data manipulation for parser improvement

Dept. of Linguistics, Indiana University

Fall 2011

Emphasizing the need for features

Zhu et al. (2011)

Source bracketing can be used as parsing constraints during decoding of a target parser

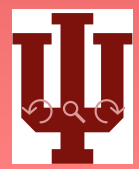
- But consider figure 1 in Zhu et al. (2011)
  - Tsinghua Chinese Treebank (TCT) tree: verb “deems” is at the right bracket of a phrase
  - Penn Chinese Treebank (CTB) tree: verb “deems” is at the left boundary of a phrase
- These annotations are *inconsistent*
  - A source parsing constraint may thus prune such a constituent

Alternative: use source bracketing structures as **guiding information**

Generic System Architecture

1. Build *source parser* & parse target treebank
2. Build a parser on (parsed-source, gold-target) sentence pairs
   - **heterogeneous parser**: incorporates information from both styles
3. Testing: take gold source parses as input and converts them

Feature-based parsing algorithms are used, to incorporate source bracketing structures



Shift-Reduce-Based Heterogeneous Parser

Shift-reduce parser uses transitions between states $< S, Q >$ (S=stack of partial parses, Q=queue of word-POS pairs)

- Heterogeneous parsing works similarly to standard way
  - Tree transformed to binary tree
  - Binary tree decomposed into gold action-state sequences
  - Train classifier on states, which are represented as feature vectors
- Feature set is augmented with features bridging current state and source parse

Features

- Target-side features: same as earlier parser
- Heterogeneous features:
  - Constituent features (e.g., bracketing matches?)
  - Relation features (e.g., nodes are identical or sisters?)
  - Frontier features (e.g., words in same base phrase?)
  - Path features (e.g., syntactic path?)

Look at table 1 to unpack these a bit ...

Experiments

Table 2 reports conversion accuracy

- All heterogeneous features improve conversion accuracy
- Impact of path feature is small, possibly due to sparseness

nb: this was done on top of POS conversion (96.2% accuracy)

Projected Treebank as Source Corpus

Jiang and Liu (2009)

Problem: Projected treebanks inherit the standard of the source

- Adapt the divergence automatically
- Boost parsing performance with additional parsed trees

Error-Tolerant Tree Projection

Many approaches directly map from source to target

- Their method works by looking for the best consistency with source trees:

$$(1)\quad \hat{T}_C = \arg\max_{T_C} C(T_C|T_E, A)$$

- Measures the degree to which Chinese tree (T_C) is consistent with English tree (T_E)
- They accumulate scores across all possible alignments, making it more error-tolerant

More details in the paper

Annotation Adaptation

Train *source parser* & parse target corpus

- Then, a *target parser* is trained
  - crucially, with guide features extracted from source parser's output

When testing, data is first parsed by source parser as an intermediate parsing result

- Then, the target parser with guide features is used
- Automatically learns the regularities of the intermediate parse

Guide Features

They work with MSTParser, tailoring guide features to it

- i.e., define features on dependency edges (cf. edge-factorization)
- Examine relationship between head and modifier in source parse:
  - Feature: Does the relationship exist? (cf. stacking)
  - Features: combinations of lexical features of MST models
- Define relationships as a variable covering these cases:
  - parent-child, child-parent, siblings, else

Experiments

Project English trees to Chinese & select those between 6 & 100 words and with a high enough projection confidence

- Source = PTB, Target = Penn Chinese Treebank (CTB)
- Source parser: 1st-order MST, Target parser: 2nd-order MST

Source parsers perform poorly (around 53%), while target parser is around 83-87% & higher than baseline

References

Jiang, Wenbin and Qun Liu (2009). Automatic Adaptation of Annotation Standards for Dependency Parsing — Using Projected Treebank as Source Corpus. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*. Paris, France: Association for Computational Linguistics, pp. 25–28.

Zhu, Muhua, Jingbo Zhu and Minghan Hu (2011). Better Automatic Treebank Conversion Using A Feature-Based Approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 715–719.