SFB 732 D5: Biased Learning for Syntactic Disambiguation

Blaubeuren - November 16, 2008
Biased Learning for Syntactic Disambiguation

- Learning from monolingual text (grammatical dependencies, n-gram language model)
- Learning from bilingual text
  - *Disambiguating ambiguous German subjects and objects* using the English translations in a German/English parallel text
  - A general approach to **improve English syntactic parsing** using the German translations in German/English parallel text
Overview
Using German to improve English parsing
D5 Contributions

Examples
Approach

Figure: English parse with high attachment (incorrect)
Overview
Using German to improve English parsing
D5 Contributions

Examples
Approach

**Figure:** English parse with low attachment (correct)
Figure: German parse with low attachment
Reranking approach using rich bitext projection features

- Goal: improve English parsing accuracy
Reranking approach using rich bitext projection features

- Goal: improve English parsing accuracy
- Working on parallel text, e.g., proceedings of European Parliament
Reranking approach using rich bitext projection features

- Goal: improve English parsing accuracy
- Working on parallel text, e.g., proceedings of European Parliament
- Begin by parsing English sentence with Bitpar (Schmid). Select 100 most probable parses
Goal: improve English parsing accuracy

Working on parallel text, e.g., proceedings of European Parliament

Begin by parsing English sentence with Bitpar (Schmid). Select 100 most probable parses

Find most probable parse of German sentence
Goal: improve English parsing accuracy

Working on parallel text, e.g., proceedings of European Parliament

Begin by parsing English sentence with Bitpar (Schmid). Select 100 most probable parses

Find most probable parse of German sentence

Using rich bitext projection features, calculate syntactic divergence of each English parse candidate and the (projection of) the German parse
Reranking approach using rich bitext projection features

- Goal: improve English parsing accuracy
- Working on parallel text, e.g., proceedings of European Parliament
- Begin by parsing English sentence with Bitpar (Schmid). Select 100 most probable parses
- Find most probable parse of German sentence
- Using rich bitext projection features, calculate syntactic divergence of each English parse candidate and the (projection of) the German parse
- Choose a high probability English parse candidate with low syntactic divergence
Rich bitext projection features

- Mix of probabilistic and heuristic features, combined in log-linear model, trained to maximize parsing accuracy
- General features: tag correspondence, span size difference, parse depth difference
- Specific features: coordination phenomena, NP structure
- Documented in EACL submission
- Current project: improve parses of Europarl corpus (1.4 million parallel sentences)
Using German to improve English parsing

D5 Contributions

D5 contributes to 3 Area D Goals, one long-term SFB goal

- Types of contextual information: D5 uses contextual information derived from bilingual and monolingual syntactic analyses, at varying levels of granularity (e.g., parse tree vs. n-gram)

Learnability of contextual information: D5 uses statistical models of context learned from bilingual and monolingual data, often itself a product of syntactic analysis

Use of contextual information: D5 uses statistical models of context for improving syntactic analysis

Incorporating linguistic insights into statistical models: D5 uses insights into complementarity of English and German ambiguity to improve statistical syntactic disambiguation
D5 contributes to 3 Area D Goals, one long-term SFB goal

- Types of contextual information: D5 uses contextual information derived from **bilingual and monolingual syntactic analyses**, at varying levels of granularity (e.g., parse tree vs. n-gram)

- Learnability of contextual information: D5 uses **statistical models of context** learned from bilingual and monolingual data, often itself a product of syntactic analysis
D5 contributes to 3 Area D Goals, one long-term SFB goal

- Types of contextual information: D5 uses contextual information derived from bilingual and monolingual syntactic analyses, at varying levels of granularity (e.g., parse tree vs. n-gram)
- Learnability of contextual information: D5 uses statistical models of context learned from bilingual and monolingual data, often itself a product of syntactic analysis
- Use of contextual information: D5 uses statistical models of context for improving syntactic analysis
D5 contributes to 3 Area D Goals, one long-term SFB goal

- Types of contextual information: D5 uses contextual information derived from bilingual and monolingual syntactic analyses, at varying levels of granularity (e.g., parse tree vs. n-gram)
- Learnability of contextual information: D5 uses statistical models of context learned from bilingual and monolingual data, often itself a product of syntactic analysis
- Use of contextual information: D5 uses statistical models of context for improving syntactic analysis
- Incorporating linguistic insights into statistical models: D5 uses insights into complementarity of English and German ambiguity to improve statistical syntactic disambiguation