Creating a semantically annotated corpus based on Discourse Representation Theory
Johan Bos  Valerio Basile  Kilian Evang  Noortje Venhuizen

The Groningen Meaning Bank (GMB)

- Current state in data-driven computational semantics:
  - Several annotated corpora are available that include some semantic annotation (PropBank, Penn Discourse TreeBank, OntoNotes)
  - However, none of these resources contain annotations that are motivated by formal semantic theory
- The objectives of the Groningen Meaning Bank are:
  - Producing a corpus of texts annotated with quasi gold-standard Discourse Representation Structures (DRSs)
  - Making this resource available for research in a kerfuffle-free manner (only public-domain texts are included)

Discourse Representation Theory (DRT)

- DRT is a theory of analysing meaning from text, in principle language-neutral
- Many linguistic phenomena are studied in the framework provided by DRT (anaphora, scope, events, tense)
- DRT has a model-theoretic backbone, allowing applications to perform inferences on the basis of first-order logic

Discourse Representation Structures (DRSs)

- DRSs are visualised as a box with two parts:
  - Top part of the box: discourse referents
  - Bottom part of the box: properties of and relations between referents
- Extensions to standard DRT:
  - neo-Davidsonian events (with VerbNet roles)
  - presuppositions (Van der Sandt, 1992)
  - rhetorical relations (Asher, 1993)

Annotation method

- Manually annotating a reasonably large corpus with gold-standard DRSs is obviously a hard and time-consuming task
- We use a bootstrapping approach that employs state-of-the-art NLP tools to get a reasonable approximation of the target annotations
- Human annotations are coming from two main sources: experts (linguists) and non-experts (players of a game with a purpose)
- The annotation of a text comprises several layers:
  - boundaries (for tokens and sentences)
  - tags (part of speech, named entities, word senses)
  - syntactic structure (based on combinatory categorial grammar)
  - semantic structure (including thematic roles and rhetorical relations)

Innovative features and possible impact of the GMB

- Comprises deep, rather than shallow semantics
  - This opens the way to empirical, data-driven approaches to computational semantics
- Integrates phenomena, instead of covering single phenomena in isolation
  - This will provide a better handle on explaining dependencies between various ambiguous linguistic phenomena
- Deals with text, not sentences
  - This gives us means to deal with ambiguities on the sentence level that require the discourse context for resolving them

Example of a DRS for a small text

```
the Australian Government allows limited fishing
```

Results

- Explorer: wiki-like interface for expert annotators
- Current corpus size (development version):
  - 70K documents, 1.3M sentences, 31M tokens
- First stable release: 1,000 documents (GMB 1.0)

The GMB Explorer: visualisation tool for manipulating DRSs

http://gmb.let.rug.nl  johan.bos@rug.nl