Extracting a verb lexicon for deep parsing from FrameNet

Mark McConville and Myroslava O. Dzikovska

Abstract: We examine the feasibility of harvesting a wide-coverage lexicon of English verbs from the FrameNet semantically annotated corpus, intended for use with a deep parser. We identify a range of non-canonical constructions for which current annotation practice leads to problems in deriving appropriate lexical entries, and discuss potential solutions.

Motivation

To use the output of a deep parser with a reasoning component, we need a wide-coverage lexicon based on a consistent set of word-senses, linked by semantic relations. Since such knowledge sources are expensive to create by hand, we are interested in extracting them from existing corpora.

FrameNet\textsuperscript{1} is a corpus of 140,000 English sentences, annotated with both syntactic and semantic information. Sentences are annotated as follows, where the target word is the verb fried:

```
Entries derived from relative clauses with overt relative pronouns: [the catfish entries derived from sentences involving a 'split' argument, e.g. Middle (e.g. Are there (Evaluate any places) Conn you want to praise Reason for their special facilities?)]
Entries derived from sentences involving a 'split' argument, e.g. [Parp the two boys] [Parp who] abducted [Victim James Bulger]
Entries derived from relative clauses with overt relative pronouns: [Parp the two boys] [Parp who] abducted [Victim James Bulger]
Entries derived from leftward extractions (e.g. Are there (Evaluate any places) Conn you want to praise Reason for their special facilities?)
```

After duplicate entries were removed from the resulting lexicon, we were left with entries for 2,600 distinct verb senses, where each verb-sense had an average of 10 distinct subcategorisation frames.

Evaluation

We hand-checked a random sample of the harvested lexical entries to verify whether they could be included in a lexicon for deep parsing. Four main sets of spurious subcategorisation frames were identified.

Subjectless frames

The harvested lexicon contained 2,200 subcategorisation frames which lack a subject argument. These were mainly derived from target verbs: (a) in the imperative; (b) with un-controlled non-overt subjects (e.g. Being accused of that is a terrible insult); and (c) with non-referential, extraposed subjects (e.g. It is to be regretted that . . .)

Passive frames

The harvested lexicon contained 4,200 subcategorisation frames which were derived from "passive" uses of a target verb, defined as follows: (a) the verb has POS-tag VBD; (b) there is a subject which invokes semantic role e; and (c) there is some other subcategorisation frame for the same verb-sense which has an object invoking role e. Using these criteria to identify passives resulted in no false positives but a large number of false negatives, many of which were caused by past participles being mistagged as VWD.

We found a further 1,000 subcategorisation frames which appeared to involve a passive argument structure with a non-VBD-tagged target verb.

As well as mistagged passives, this picked up the following "passive-like" uses of target verbs:

- those in the complement of raising adjectives ([Ext Planning and control] are difficult to achieve in this form of production)
- those in the need-ing construction ([Ext Many private problems] need airing in the family)
- middles ([Ext You] frighten easily)

Of the frames identified as "passive" by the weaker criteria, 220 turn out to be false positives. Most of these involve target verbs in the causative-inchoative alternation (e.g. John's back arched vs. John arched his back). Although FrameNet policy is to distinguish between inchoative (e.g. Motion) and causative (Causative Motion) frames, linked by the Causative relation, this has not been applied rigorously in practice.

Frames containing modifiers

Many of the remaining subcategorisation frames include at least one argument which must linguists would adjudge to be a "modifier" or "adjunct". The FrameNet ontology distinguishes between Core and non-Core semantic roles, providing a useful approximation to the argument-adjunct distinction. However, the correspondence is not exact, since there are many examples of direct objects assigned non-Core roles, e.g. I ripped [Subregion the top] from my packet of cigarettes.

Frames involving control

FrameNet annotation does not identify the controller of understood subjects of VPto arguments. We attempted to distinguish between subject and object control by looking for sentences that have been independently annotated for both the control verb and the head of its controlled argument. Unfortunately, the data proved to be too sparse for this to be successful.

Conclusion

After deleting subjectless frames, passive frames and modifiers, the harvested lexicon contained 9,000 subcategorisation frames for 2,600 verb-senses, giving an average of 3.4 frames per verb-sense. Currently, most of the entries have been syntactically aligned with the TRIPS domain-independent dialogue grammar\textsuperscript{2}, allowing frame-semantic representations to be constructed using the TRIPS parser.

\textsuperscript{1}http://framenet.icsi.berkeley.edu
\textsuperscript{2}http://www.cs.rochester.edu/research/trips