PropBank Unification Documentation

(July 2015 - by Tim O'Gorman)

The new Unified (AMR-style) PropBank rolesets:

Prior PropBank frame files, and the rolesets (individual senses) within them, were subtyped by part of speech tags, and gave various senses for a single lemma. Earlier AMR releases utilized the verbal subset of those frames, and had annotators generalize across parts of speech to the nearest verbal sense. The unified PropBank rolesets, in order to add power and coverage to the Abstract Meaning Representation project, essentially formalize that process by lumping etymologically related lemmas with the same sense into the same roleset. These rolesets, in AMR style, are now closer to a representation of an actual concept, and represent the different forms that the roleset might take with a new "alias" element that expresses its possible realizations.

Alias field conventions and assumptions:

- Aliases represent a set of etymologically related lemmas that might be used to refer to that particular roleset.
- All aliases have a lemma and a correponding part of speech tag (the "pos") field, with "v", "n" and "j" representing verbs, nouns and adjectives, respectively.
- Multi-word aliases are separated by an underscore, "_", mostly being verb+particle constructions such as "throw_up".
- Common patterns of how verbs combine with prepositions, objects or adverbs are not currently represented in the alias fields. This means that the sense "have yet to" (have.11) is aliased merely by the verbal alias "have".
- Participles (-ed and -ing forms) will appear as adjectival or nominal aliases only when commonly POS-tagged as such during treebanking.
- All aliases are lemmatized as appropriate for their Treebank part of speech tags, so that adjectives ending in "-ed" (such as "spirited") or nouns ending in "-ing" (such as "accounting") do not have those suffixes removed.

Roleset naming conventions and assumptions:

- Since there are many possible lemmas for each roleset, one can no longer assume that the roleset name and the lemma in the sentence are identical, as one could with prior PropBank releases.
- A roleset is named after the most common verbal alias (if one exists), backing off to nominal and then adjectival aliases if no verbal alias is present.
- Rolesets are always named after one of the aliases, and may be named after verb+particle aliases (as in "throw_up.05")
- Roleset IDs are numbered so that each sense number occurs only once for each frame file. "criticism.04" does not entail that there are three other "criticism" rolesets, but simply that it is the fourth roleset within the frame file "criticize.xml", alongside criticize.01, critical.02 and critical.03.
- Roleset names default to American spelling conventions whenever there are multiple spelling

variants.

Factors Determining Which Old Rolesets were Merged

- When combining different parts of speech into the new unified rolesets, senses were merged only if they were etymologically related and sufficiently similar in meaning.
- FrameNet and VerbNet were consulted in making these judgements of similarity.
- Rolesets were only merged when they shared the same arguments.
- Some aspectual distinctions were merged into singular frames, but strong differences, such as the difference between causative verbs and states (as in "blacken" and "black") were left separate.

PropBank Frame File conventions:

- Rolesets are clustered into frame files with the intent that relatively simple heuristics should be able to find the frame file containing the correct sense.
- This means that frame files were merged if they contained any overlapping aliases. For example, rolesets for "flee" and "fly" are now both in contained within "fly.xml", so that one might get to the correct frame file when observing the word "flight".