

Introduction

- ▶ The Proposition Bank project: additional layer of predicate-argument information, or semantic role labels, on top of the syntactic structures of the **Penn Treebank**. The Proposition Bank assigns semantic roles to nodes in the syntactic trees of the Penn Treebank.
- ▶ The resulting resource is **shallow** in that it does not represent coreference, quantification, and many other higher-order phenomena
- ▶ At the same time it is also **broad**, in that it covers every instance of every verb in the corpus and allows representative statistics to be calculated.
- ▶ PropBank annotates verbs; the **NomBank** sister project annotates nouns.
- ▶ Intended from the get-go as a **resource for training statistical role-semantic parsers**.

PropBank annotations

- ▶ **RoleSet**: A set of roles corresponding to a distinct usage of a verb is called a roleset, and can be associated with a set of syntactic frames indicating allowable syntactic variations in the expression of that set of roles. The roleset with its associated frames is called a **Frameset**.
 - ▶ PB annotates some adjuncts in addition to arguments
 - ▶ ARG[0-9] are defined on verb-by-verb basis
 - ▶ **ARG0**: typically something like a **proto-Agent**
 - ▶ **ARG1**: typically something like a **proto-Patient**
 - ▶ No consistent generalizations can be made across verbs for the higher numbered arguments
 - ▶ Effort was made to consistently define roles across members of VerbNet classes.
 - ▶ ARGM-roles are taken not to be verb-specific

More on PB annotations

- ▶ Arg-numbering: to be **theory-neutral**
- ▶ Usu 2-4 ARGs, sometimes as many as 6
- ▶ Types of ARGM

LOC: location

EXT*: extent

DIS: discourse connectives

ADV: general-purpose

NEG: negation marker

MOD: modal verb

- ▶ other secondary tags: PRD

CAU: cause

TMP: time

PNC: purpose

MNR: manner

DIR: direction

Yet more on PropBank annotations

- ▶ A polysemous verb may have more than one Frameset, when the differences in meaning are **distinct enough**.
- ▶ Syntactic-semantic criteria go into this
- ▶ Alternations which preserve verb meanings, such as causative/inchoative or object deletion are considered to be one frameset only.
- ▶ Verb-particle combinations are always distinct framesets
- ▶ Some differences to FN
 - ▶ Symmetric-asymmetric construal alternations are not explicitly marked by different role labels (we met; I met him)
 - ▶ No account of omitted arguments

Even More on PB annotations

- ▶ Standoff format that references nodes in Penn Treebank
 - ▶ wsj/00/wsj_0083.mrg 16 9 acceleration 01 9:0-rel 10:0,11:1-ARG1
 - ▶ wsj/01/wsj_0115.mrg 2 24 acceleration 01 24:0-rel 25:1-ARG1
 - ▶ ...
- ▶ The framesets can be viewed as extremely **coarse-grained sense distinctions**, with each frameset corresponding to one or more of the Senseval 2 WordNet 1.7 verb groupings. Each grouping in turn corresponds to several WordNet 1.7 senses.
- ▶ Each instance of a polysemous verb is marked as to which frameset it belongs to, with **inter-annotator agreement of 94%**.

Practicalities of PB

- ▶ Annotators are presented with the roleset descriptions and the syntactic tree.

Practicalities of PB

- ▶ Annotators are presented with the roleset descriptions and the syntactic tree.
- ▶ They mark the appropriate nodes in the tree with role labels.
- ▶ The lexical heads of constituents are not explicitly marked either in the Treebank trees or in the semantic labeling.
- ▶ Labelers cannot modify the syntax; they can label more than one node.

Practicalities of PB

- ▶ Annotators are presented with the roleset descriptions and the syntactic tree.
- ▶ They mark the appropriate nodes in the tree with role labels.
- ▶ The lexical heads of constituents are not explicitly marked either in the Treebank trees or in the semantic labeling.
- ▶ Labelers cannot modify the syntax; they can label more than one node.
- ▶ PP Args are treated differently than PP Adjuncts
[Arg1 Its net income] **declining** [Arg2-EXT 42%] to [Arg4 \$121 million] [ArgM-TMP in the first 9 months of 1989]. (wsj 0067)

Practicalities of PB

- ▶ Annotators are presented with the roleset descriptions and the syntactic tree.
- ▶ They mark the appropriate nodes in the tree with role labels.
- ▶ The lexical heads of constituents are not explicitly marked either in the Treebank trees or in the semantic labeling.
- ▶ Labelers cannot modify the syntax; they can label more than one node.
- ▶ PP Args are treated differently than PP Adjuncts
[Arg1 Its net income] **declining** [Arg2-EXT 42%] to [Arg4 \$121 million] [ArgM-TMP in the first 9 months of 1989]. (wsj 0067)
- ▶ Annotation of traces
[Arg0 John_i] tried [Arg0 trace_i] to **kick** [Arg1 the football], but Mary pulled it away at the last moment.

PB Workflow

- ▶ Word sense division/grouping
- ▶ Care taken to make synonymous verbs (usu in the sense of sharing a verbnet class) have the same framing, with the same number of roles and the same descriptors on those roles.

PB Workflow

- ▶ Word sense division/grouping
- ▶ Care taken to make synonymous verbs (usu in the sense of sharing a verbnet class) have the same framing, with the same number of roles and the same descriptors on those roles.
- ▶ Generally, a given lexeme/sense pair required about 10-15 minutes to frame

PB Workflow

- ▶ Word sense division/grouping
- ▶ Care taken to make synonymous verbs (usu in the sense of sharing a verbnet class) have the same framing, with the same number of roles and the same descriptors on those roles.
- ▶ Generally, a given lexeme/sense pair required about 10-15 minutes to frame
- ▶ Annotation is a two-pass, blind procedure followed by adjudication
- ▶ Both role labeling and the choice of frameset are adjudicated

Inter-annotator Agreement

		P(A)	P(E)	κ
including ArgM	role identification	.99	.89	.93
	role classification	.95	.27	.93
	combined decision	.99	.88	.91
excluding ArgM	role identification	.99	.91	.94
	role classification	.98	.41	.96
	combined decision	.99	.91	.93

Example Frameset

- ▶ Frameset accept.01 “take willingly”
- ▶ Arg0: Acceptor
- ▶ Arg1: Thing accepted
- ▶ Arg2: Accepted-from
- ▶ Arg3: Attribute
- ▶ Ex:[Arg0 He] [ArgM-MOD would][ArgM-NEG n't] accept [Arg1 anything of value] [Arg2 from those he was writing about]. (wsj 0186)

Historical Context: NLP

- ▶ While the Penn Treebank provides semantic function tags such as temporal and locative for certain constituents (generally syntactic adjuncts), it does not distinguish the different roles played by a verb's grammatical subject or object in the above examples.
- ▶ PropBank's semantic role annotation process begins with a rule-based automatic tagger, the output of which is then hand-corrected
- ▶ Pre-PropBank, information extraction systems relied on a shallower level of semantic representation, similar to the level adopted for the Proposition Bank, but they tended to be very domain specific.
- ▶ The systems were trained and evaluated on corpora annotated for semantic relations pertaining to, for example, corporate acquisitions or terrorist events.

Historical context: Alternation studies: Levin 1993

- ▶ Groups verbs into classes based on shared syntactic behavior
- ▶ Assumption: syntax reflects semantics, in particular components of meanings
- ▶ Hot issue: how regular/strong/reliable is the connection?
- ▶ **VerbNet** extends Levin's classes by adding an abstract representation of the syntactic frames for each class with explicit correspondences between syntactic positions and the semantic roles they express (e.g. Agent REL Patient, or Patient REL into pieces for break)

Historical context: Alternation studies II

- ▶ Objective of Proposition Bank is not a theoretical account of how and why syntactic alternation takes place, but rather to provide a useful level of representation and a corpus of annotated data to enable empirical study of these issues.
- ▶ There is only a 50% overlap between verbs in VerbNet and those in the Penn TreeBank II
- ▶ PropBank itself does not define a set of classes, nor does it attempt to formalize the semantics of the roles it defines.
- ▶ Lexical resources such as Levin's classes and VerbNet provide information about alternation patterns and their semantics, but the frequency of these alternations and their effect on language understanding systems has never been carefully quantified.

Historical context: Alternation studies III

- ▶ While learning syntactic subcategorization frames from corpora has been shown to be possible with reasonable accuracy , such work usually does not address the semantic roles associated with the syntactic arguments.
- ▶ More recent work has attempted to group verbs into classes based on alternations, usually taking Levin's classes as a gold standard
- ▶ But without an annotated corpus of semantic roles, this line of research has not been able to measure the frequency of alternations directly, or, more generally, to ascertain how well the classes defined by Levin correspond to real world data.

References

- ▶ Martha Palmer, Dan Gildea, Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics.
- ▶ Levin, Beth. 1993. English Verb Classes And Alternations: A Preliminary Investigation. University of Chicago Press, Chicago.
- ▶ Kipper, Karin, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-2000), Austin, TX, July-August.