

Corpus Linguistics (L615)

Semantic annotation

Markus Dickinson

Department of Linguistics, Indiana University
Spring 2009

Corpora with semantic annotation are increasingly relevant in natural language processing

- ▶ See: Baker et al. (1998); Palmer et al. (2005); Burchardt et al. (2006); Taulé et al. (2005)

Semantic role labeling

- ▶ used for tasks such as:
 - ▶ information extraction (Surdeanu et al. 2003)
 - ▶ machine translation (Komachi et al. 2006)
 - ▶ question answering (Narayanan and Harabagiu 2004)
- ▶ requires corpora annotated with predicate-argument structure for training and testing data
 - ▶ Gildea and Jurafsky (2002); Xue and Palmer (2004); Toutanova et al. (2005); Pradhan et al. (2005), ...

Semantically-annotated corpora also have potential as sources of linguistic data for theoretical research

[Semantic annotation](#)[Propbank](#)[Other English resources](#)[SALSA](#)[References](#)

Need feedback on annotation schemes:

- ▶ difficult to select an underlying theory (see, e.g., Burchardt et al. 2006)
- ▶ difficult to determine certain relations, e.g., modifiers (ArgM) in PropBank (Palmer et al. 2005)
- ▶ Not a clear consensus on what elements to tag and how to tag them (Palmer et al. 2000)

Broadly speaking, there are 2 main ways to do semantic annotation:

- ▶ Lexical semantics: word senses
 - ▶ The major issue here is how to deal with polysemy
 - ▶ How many senses does each word have, and what are they?
- ▶ Compositional semantics: argument relations
 - ▶ The connection to syntax is apparent
 - ▶ Requires an inventory of argument roles & relies on the particular verb sense

Work by Beth Levin and others shows that these concepts are interrelated in some ways

Sense Tagging the Penn Treebank

Palmer et al. (2000)

Corpus Linguistics

Semantic
annotation

Semantic
annotation

Propbank

Other English
resources

SALSA

References

Initially tagged a 5000-word corpus (later expanded for Propbank)

- ▶ Selected WSJ articles which contained “interesting verbs” & covered a range of topics
- ▶ Sense-tagged only the verbs & headwords of arguments/adjuncts
- ▶ Used WordNet senses
 - ▶ Additionally tagged proper nouns as person, company, date, or name

Sense-tagged 2100 words, with inter-annotator agreement rate of 89%

- ▶ 700 verbs: 81% agreement
- ▶ 350 verb lemmas: 90 had at least one disagreement occurrence
 - ▶ WordNet did not have correct sense, or
 - ▶ WordNet did not adequately define the senses

Sample sense-tagged text

Corpus Linguistics

Semantic
annotation

Semantic
annotation

Propbank

Other English
resources

SALSA

References

```
<wf lemma=Donald_Trump wnsn=person>Donald Trump</wf> ,  
<wf lemma=who wnsn=person:DT>who</wf>  
<wf cmd=arb lemma=face wnsn=?>faced</wf> rising  
<wf cmd=done lemma=doubt wnsn=1>doubt</wf> about his  
<wf cmd=done lemma=bid wnsn=2>bid</wf> for American Airlines  
<wf lemma=AMR_Corp. wnsn=company>AMR Corp.</wf> even before  
...
```

Predicate-argument structure

Building from the sense annotation, they also annotated predicate-argument structure

- ▶ Added subscripts to PTB trees to indicate what semantic role a constituent plays in a sentence
 - ▶ e.g., SBJ on an NP indicates a subject role
 - ▶ e.g., TMP on a PP indicates temporal information about an event

To obtain predicate-argument annotation, verbs needed to be linked to their arguments

- ▶ Required being able to automatically determine semantic heads of phrases
- ▶ Morphological information & phrasal lexical entries were also added

Predicate-argument relations (formally)

Semantic annotation is non-uniform:

(1) [*Arg*₁ lending practices] **vary**/vary.01 [*Arg*_{2-EXT} widely]
[*Arg*_{M-MNR} by location]

1. the verb sense
2. the span of each argument
3. argument label names

Predicate-argument & verb sense annotation are really 2 different things (cf. semantic role labeling, Morante and van den Bosch 2007)

Criteria for senses

shake mismatches

Corpus Linguistics

Semantic
annotation

Semantic
annotation

Propbank

Other English
resources

SALSA

References

Data trained with one sense tagset evaluated on a different tagset

- ▶ Need to figure out mismatches

For *shake*:

- ▶ WordNet 1.6: 8 senses + 5 for *shake up* & 2 for *shake off*
- ▶ Hector: 8 senses + 3 for *shake up*, 2 for *shake off*, 3 for *shake down*, & 2 for *shake out*

There are fundamental differences in the organization, however

- ▶ Hector distinguishes:
 - ▶ *shaking hands with someone*
 - ▶ *shaking one's fist*
 - ▶ *shaking one's head*
- ▶ Hector distinguishes:
 - ▶ intransitive TREMBLE sense (*My hands were shaking from the cold*)
 - ▶ proactive MOVE sense (*He shook the bag violently*)
- ▶ These last 2 + hand-shaking are grouped into WordNet sense WN1

WordNet distinguishes the type of action

- ▶ WN2: gentle tremors
- ▶ WN3: rapid vibrations
- ▶ WN4: swaying

Hector1 maps equally onto WN1, WN2, WN3, & WN4

Criteria for senses

Concrete criteria

To determine how to split up a word like *shake* into different senses, different criteria are used:

- ▶ Different specific lexical items
- ▶ Different syntactic frames
- ▶ Different semantic class constraints (or preferences) on arguments
- ▶ Different outcomes

Criteria for senses

Major divisions of *shake*

The 27 Hector senses & 15 WordNet senses for *shake* can be grouped into 5 major divisions:

- ▶ Externally controlled motion (causative, inchoative, resultative)
 - ▶ includes idioms
 - ▶ can be amplified with prepositions (*up*, *down*, etc.)
- ▶ Funnel, involving change of location (causative)
- ▶ Body-internal state, cf. trembling (causative, inchoative)
- ▶ Crane, involving particular body parts (causative, cognate object)
- ▶ Psych/amuse (causative, no middle)

Some consistency issues

Dickinson and Lee (2008)

It is hard to maintain predicate-argument consistency, especially when built on top of other layers of annotation:

- (2) a. coming/VBG [_{Arg1} months] ,
b. coming/JJ months ,
- (3) a. [_{Arg1} net income in its first half] rose 59 %
b. [_{Arg1} net income] in its first half rose 8.9 %
- (4) a. That could [_{Arg2-MNR} substantially] reduce the value of the television assets .
b. the proposed acquisition could [_{ArgM-MNR} substantially] reduce competition ...

- ▶ Some verbs are ambiguous in whether they take arguments and what type of arguments they take

(5) a. [_{Arg1} Analysts] **had** mixed responses

b. [_{Arg1} Analysts] had **expected** Consolidated to post a slim profit ...

- ▶ Much argument identification ambiguity rooted in difficulties resolving syntactic ambiguity

(6) a. **seeking** [_{Arg1} a buyer] [_{PP} for several months]

b. **seeking** [_{Arg1} a buyer for only its shares]

- ▶ Some argument relations depend upon the sense of the verb, which depends upon other arguments of verb

(7) a. [_{Arg0} he] will **return** Kidder to prominence

b. [_{Arg1} he] will **return** to his old bench

Other English resources

SemCor

Texts semantically annotated with WordNet 1.6 senses
(Rada Mihalcea)

- ▶ With links to later WordNet sense inventories (1.7, ..., 3.0)

<http://www.cse.unt.edu/~rada/downloads.html#semcor>

From README:

brown1	103 semantically tagged Brown Corpus files (all content words tagged)
brown2	83 semantically tagged Brown Corpus files (all content words tagged)
brownv	166 semantically tagged Brown Corpus files (only verbs tagged)

Corpus Linguistics

Semantic
annotation

Semantic
annotation

Propbank

Other English
resources

SALSA

References

Other English resources

DSO

The DSO corpus (Hwee Tou Ng and Hian Beng Lee) has about 192,800 words tagged with WordNet senses

- ▶ 121 nouns & 70 verbs: among the most frequently occurring
- ▶ Available from the LDC: <http://www ldc upenn edu / Catalog / CatalogEntry.jsp?catalogId=LDC97T12>

Example:

```
ca01.db #020 ‘‘ These >> actions 8 << should serve to protect  
fact and in effect the court 's wards from undue costs and  
appointed and elected servants from unmeritorious criticism  
the jury said .
```

Corpus Linguistics

Semantic
annotation

Semantic
annotation

Propbank

Other English
resources

SALSA

References

Other English resources

FrameNet

FrameNet is an online lexical resource for English (with FrameNets also in other languages)

- ▶ <http://framenet.icsi.berkeley.edu/>

FrameNet features

- ▶ 10,000 lexical units, with more than 825 semantic frames
- ▶ 135,000 annotated sentences

We'll talk more about frame semantics momentarily ...

The SALSA project

Burchardt et al. (2006)

Large corpora and large domain-independent lexica can help the study of:

- ▶ lexical semantics
- ▶ syntax-semantics linking properties
- ▶ noncompositional phenomena, e.g., idiomatic & metaphoric expressions
- ▶ cross-lingual analysis & application of lexical semantic information
 - ▶ particularly apt for frame semantics, as it has a common, largely language-independent word sense & role inventory

Project page: <http://www.coli.uni-saarland.de/projects/salsa/>

See <http://www.coli.uni-saarland.de/projects/salsa/corpus/>
for the release

Corpus Linguistics

Semantic
annotation

Semantic
annotation

Propbank

Other English
resources

SALSA

References

Frame semantics describes meaning as:

- ▶ characterized by the background knowledge necessary to understand each expression
- ▶ A *frame* is evoked by a word or expression
 - ▶ Coarse-grained frame descriptions generalize over different lexical items (unlike Propbank)
- ▶ Each frame has its own set of semantic roles, called *frame elements*
 - ▶ Participants & propositions of an abstract situation
 - ▶ Frame elements are local to individual frames, instead of using universal roles

Frame example: STATEMENT

Frame description

This frame contains verbs and nouns that communicate the act of a SPEAKER to address a MESSAGE to some ADDRESSEE using language. A number of the words can be used performatively, such as declare and insist.

Corpus Linguistics

Semantic
annotation

Semantic
annotation

Propbank

Other English
resources

SALSA

References

Frame example: STATEMENT

Frame elements

- ▶ SPEAKER: **Evelyn** said she wanted to leave.
- ▶ MESSAGE: Evelyn announced **that she wanted to leave.**
- ▶ ADDRESSEE: Evelyn spoke **to me** about her past.
- ▶ TOPIC: Evelyn's statement **about her past.**
- ▶ MEDIUM: Evelyn preached to me **over the phone.**

Frame example: STATEMENT

Predicates

- ▶ acknowledge.v
- ▶ acknowledgement.n
- ▶ add.v
- ▶ address.v
- ▶ admission.n
- ▶ admit.v
- ▶ affirm.v
- ▶ affirmation.n
- ▶ allegation.n
- ▶ ...

Syntax-semantics examples

Frame semantics is between syntax and “deep” semantics

e.g., generalizes over verbal alternations:

- ▶ [Peter]_{agent} hit_{cause; impact} [the ball]_{impactee}.
- ▶ [The ball]_{impactee} was hit_{cause; impact}.

and over nominalizations:

- ▶ [Evelyn]_{speaker} spoke_{statement} [about her past]_{topic}.
- ▶ [Evelyn's]_{speaker} statement_{statement} [about her past]_{topic}

Annotation for German

Build on top of the TIGER corpus of German

- ▶ Single flat tree for each frame
- ▶ Root node labeled with frame name; edges with frame element names
 - ▶ Frame elements refer to syntactic constituents

Communication_response:

(8) “ [S **Schlecht**]_{Message} ”, antworter [NP
“ Badly ”, answers the
die Branche]_{Speaker} [PP im Chor] .
industry section in unison .

Annotation proceeds one predicate at a time & all instances of a predicate are annotated

Issues which arise in extending FrameNet to German

- ▶ Cross-lingual divergences, e.g., for ASSISTANCE:
 - ▶ FOCAL_ENTITY (*with*-PP) and GOAL (VP/S) FEs can be distinguished in English by syntactic criteria
 - ▶ But German has *bei*-PPs with deverbal nouns, causing confusion
- ▶ Missing frame elements, e.g., dative objects
 - ▶ TAKING: the SOURCE can be either a possessor or location in English, but both not allowed in same sentence
 - ▶ German: both are allowed → add POSSESSOR FE
- ▶ Differences in lexical realization patterns
 - ▶ *fahren* could mean *drive* (OPERATE_VEHICLE) or *ride* (RIDE_VEHICLE)
 - ▶ Have to underspecify frame for these cases

Extending to German (cont.)

For each word instance, check whether FrameNet frame applies

- ▶ For non-covered readings, group instances into “sense groups”
- ▶ For each group, create a predicate-specific proto-frame

For 476 German predicates: 18,500 instances with 628 frames (2.8/predicate)

- ▶ 252 FrameNet frames (2.0/predicate)
- ▶ 373 new proto-frames (0.8/predicate)

- ▶ Support Verb Constructions
- ▶ Idioms: annotate complete multiword unit as frame-evoking element
- ▶ Metaphors, e.g., *unter die Lupe nehmen*: *to put* (lit. *take*) *under a magnifying glass*:
 - ▶ Source frame models syntactic realization patterns (e.g., TAKING)
 - ▶ Target frame models the understood meaning (e.g., SCRUTINY)
- ▶ Vagueness: annotators can assign more than 1 label (for frames or frame elements)

References

- Baker, Collin F., Charles J. Fillmore and John B. Lowe (1998). The Berkeley FrameNet Project. In *Proceedings of ACL-98*. Montreal, pp. 86–90.
- Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado and Manfred Pinkal (2006). The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of LREC-06*. Genoa.
- Dickinson, Markus and Chong Min Lee (2008). Detecting Errors in Semantic Annotation. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*. Marrakech, Morocco.
- Gildea, Daniel and Daniel Jurafsky (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics* 28(4), 245–288.
- Komachi, Mamoru, Masaaki Nagata and Yuji Matsumoto (2006). Phrase Reordering for Statistical Machine Translation Based on Predicate-Argument Structure. In *Proceedings of the International Workshop on Spoken Language Translation*. Kyoto, Japan, pp. 77–82.
- Morante, Roser and Antal van den Bosch (2007). Memory-Based Semantic Role Labeling of Catalan and Spanish. In *Proceedings of RANLP-07*. pp. 388–394.
- Narayanan, Sridhar and Sanda Harabagiu (2004). Question Answering based on Semantic Structures. In *International Conference on Computational Linguistics (COLING 2004)*. Geneva, Switzerland.
- Palmer, Martha, Hoa Trang Dang and Joseph Rosenzweig (2000). Sense Tagging the Penn Treebank. In *Proceedings of the Second Language Resources and Evaluation Conference, LREC-00*. Athens.
<http://verbs.colorado.edu/~mpalmer/papers/lrec00.ps.gz>

- Palmer, Martha, Daniel Gildea and Paul Kingsbury (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 31(1), 71–105.
- Pradhan, Sameer, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin and Daniel Jurafsky (2005). Support Vector Learning for Semantic Argument Classification. *Machine Learning* 60(1), 11–39.
- Surdeanu, Mihai, Sanda Harabagiu, John Williams and Paul Aarseth (2003). Using Predicate-Argument Structures for Information Extraction. In *Proceedings of ACL-03*.
- Taulé, M., J. Aparicio, J. Castellví and M.A. Martí (2005). Mapping syntactic functions into semantic roles. In *Proceedings of TLT-05*. Barcelona.
- Toutanova, Kristina, Aria Haghighi and Christopher Manning (2005). Joint Learning Improves Semantic Role Labeling. In *Proceedings of ACL-05*. Ann Arbor, Michigan, pp. 589–596.
- Xue, Nianwen and Martha Palmer (2004). Calibrating Features for Semantic Role Labeling. In Dekang Lin and Dekai Wu (eds.), *Proceedings of EMNLP 2004*. Barcelona, pp. 88–94.