

Lexical Semantics, Semantic Roles and Semantic Role Labeling

Adam Meyers
New York University



Outline

- Basics of Lexical Semantics
- Word Senses and WordNet
- Semantic Roles
- Semantic Role Labeling
- Selection Restrictions



Lemmas and Wordforms

- Lemma: basic word form (paired with POS) used in lexicon
 - base form representing a set of inflected forms
 - Singular nouns: **book** → **book, books**
 - Bare infinitive verbs: **be** → **be, being, been, am, is, are, were, was**
 - Base adjective: **angry** → **angry, angrier, angriest**
- Word form: word how it actually occurs
 - a single wordform can be related to multiple lemmas:
 - **bases** is the plural of **basis** and **base**
 - **leaves** is the plural of **leaf** and **leave**
 - A word form can be defined phonologically (different homophones)
 - /**tu**/ has 3 lemmas corresponding to **two, to** and **too**
 - A word form can be defined orthographically (different homographs)
 - Thus **does** corresponds to 2 lemmas with different pronunciations: (1) present 3rd person singular of the verb **do** and (2) plural of the noun **doe**



Senses

- Conventional Dictionaries and Thesauri map lemmas to sets of different meanings called senses
- Granularity of senses: a standard problem in lexicography
 - **merge** 2 senses together or **split** one sense into 2?
- Lets lookup **bank** in WordNet (Version 3), a thesaurus/dictionary that we will be featuring
 - Compare definitions 1 and 2
 - Compare definitions 2 and 9
 - Is it possible that these definitions should be merged together?
 - All organization names can stand in for buildings that house them
 - Thus 9 is predictable from 2 (by metonymy)
 - A single instance can have both “senses” at once:
 - *The bank on the corner hired 3 security guards*

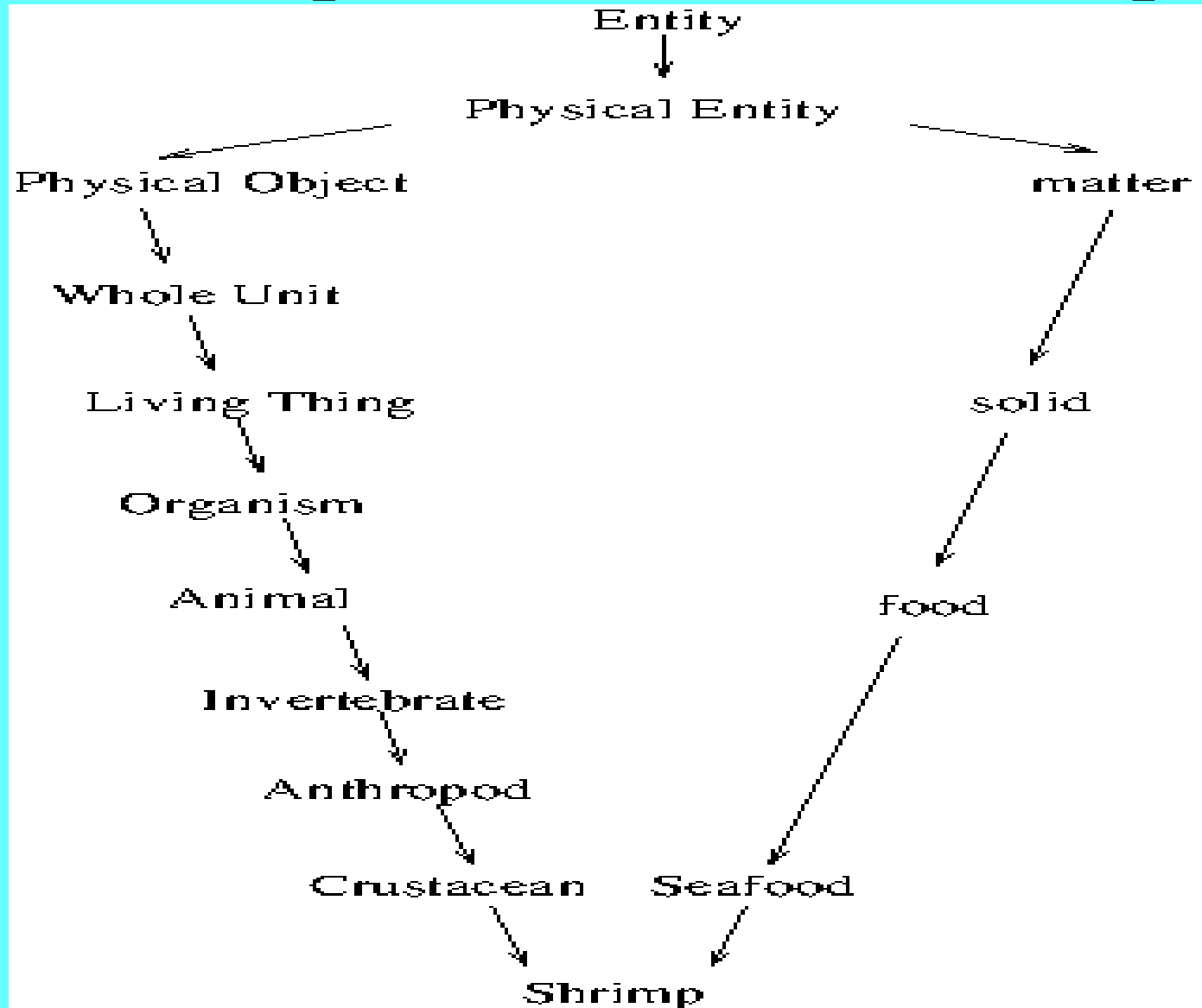


WordNet Senses

- Conventional Dictionaries
 - Senses are defined informally as text
 - The relations between senses is not represented
- WordNet
 - a sense is defined by a *Synset*, the set of words that share that meaning. Intuition = define properties by extensions
 - Formal semantic definitions are often *extensional*
 - Both Frege and Russel's set-theoretic definitions of natural numbers
 - a number N is defined as the set of sets with cardinality N
 - Many other definitions of meaning used in (computational) linguistics
 - Senses are hierarchical
 - Senses have hyponyms (sub-senses) and hypernyms (super-senses)
 - *furniture* \supset *seat* \supset *chair* \supset *recliner*
 - The full graph allows multiple inheritance
 - and even cycles such as : *restrain* \supset *inhibit* \supset *restrain* \supset *inhibit* ...



The Super Classes of *shrimp*



Some CL uses for WordNet

- Sense disambiguation (multi-sense words only):
 - Given some text tagged with WN senses
 - Train a classifier that can automatically tag raw data
 - Issue: fine-grained senses are difficult to tag (manually or automatically)
 - Many approaches collapse several WordNet senses together
- Calculating Semantic Similarity
 - Some models of similarity of meaning are based on distance in the sense hierarchy
 - Issue: same length paths do not reflect same similarity
 - distinctions are based on available information, not carefully calculated similarity distances
 - cruder relative path distances are used or path distances are combined with other information, e.g., similarity in text (in terms of n-grams)



Extensional Definition of Sentence Meaning

- A Sentence Meaning is defined by the set of sentences that share that meaning
- Problem: 2 sentences rarely mean exactly the same thing
 - *The doctor examined the patient* \approx *The patient was examined by the doctor*
 - Different focus, sometimes effects interpretation in context
 - Add the adverb *willingly* before *examined*
 - *I convinced the doctor to examine the patient* \neq *I convinced the patient to be examined by the doctor*
- Fudge: Weaker requirement than synonymy
 - Truth Value: $A = B$ iff A and B are both True or both False
 - Define relative to some task or domain
 - *He was dead (at the time)* = *He wasn't alive (at the time)*
 - If we do not consider those not yet born.



Semantic Roles Before 1965

1. *The tourists* were eaten by *the clam* ↔ *The clam* ate *the tourists*
2. *The clam* that ate *the tourists* is at NYU → *The clam* ate *the tourists*
3. *The clam* succeeded in eating *the tourists* → *The clam* ate *the tourists*
4. *The tourist* broke *the clam's shell* → *The clam's shell* broke
5. *John* ate *chicken* and *Mary* avocado → *John* ate *chicken* & *Mary* ate *avocado*

ETC

- For thousands of years, linguists described lexically and syntactically predictable paraphrase. Several alternative approaches are in use today.
- About 2500 years ago, Panini assigned verb/argument relations (karakas) corresponding to the English terms:
 - source, beneficiary, instrument, location, patient, agent
- In these terms, *the clam* is in an **agent** relation with *eat* and *the tourists* is in a **patient** relation in 1 to 3, whereas *the tourist* is the **agent** and *the clam's shell* is the **patient** in 4. These roles are described independently of where the phrases appear.
- These (and similar concepts) are the precursors to today's semantic roles (aka case roles, theta roles, thematic roles), starting with a 1965 Dissertation by J. Gruber




Semantic Roles are Relational

- A Role is a relation between a predicate and an argument
 - *The clam*/AGENT *ate* *the tourists*/PATIENT
 - *AGENT(ate, The clam)*
 - *PATIENT(ate, the tourist)*
- Contrast with POS & constituent tags which label individual units
- Usage:
 - ARG is the ROLE (of the PRED)
 - *The clam* is the agent (of *ate*)
 - *The tourist* is the patient (of *ate*)
 - PRED assigns ROLE to ARG
 - *ate* assigns *The clam* the role **agent** and *the tourist* the role **patient**
- Thus a single constituent can bare multiple roles
 - *They ate the clam that ate the tourists*
 - *the clam* is the patient of one instance of *ate* and the agent of the other instance



Empty Categories and Shared Structure

- Harris, Chomsky, etc. use Empty Categories to represent canonical word order
 - Map word order to Tense Indicative Clause (or subset)
 - *The clam₂ e₁ was eaten ₋₂ by the tourist₁*
 - *The tourist₁ ate the clam₂*
 - *The tourist wanted ₋₁ to eat the clam₂*
- Other Represent Semantic Relations in Graphs
 - *The clam wanted to be eaten by the tourist*

 - These edges can be labeled:
 - subj, obj, ind_obj (or 1, 2, 3, ... or Arg0, Arg1, Arg2, ...)
 - agent, theme, patient, ...



Fillers and Gaps

- Sometimes NLP discuss these constructions independently of how they are represented as filler/gap constructions
 - Filler = “Antecedent” of “missing” phrase
 - Gap = canonical position, where the phrase occurs in declarative non-passive sentence with no missing arguments
- Relation between Filler and Gap
 - Usually, represent the exact same object/action/etc in the world
 - Mary₁ wants ___₁ to be a linguist
 - Sometimes the filler and gap are different tokens of the same type
 - Mary ate ___₁ and John did not eat ___₁ a sandwich₁
 - Sometimes a gap is filled by an object that is not a constituent in the sentence
 - Mary₁ talked to John₂ ___₁₊₂ about leaving the party together.
- A descriptively adequate account needs to handle all cases.



Semantic Role Labels Today

- Semantic Roles 1960s and 1970s:
 - ***The Gorilla**/Agent bit **the zoo keeper**/Patient*
 - ***The train**/THEME traveled from **Boston**/SOURCE to **New York**/GOAL*
 - ***Bad television shows**/THEME annoy **consumers**/EXPERIENCER*
- Problem: it has proven impossible to define an inventory of roles that is both clearly defined and exhaustive
 - Solution 1: Assume that roles are specific to each verb (e.g., number them)
 - Relational Grammar of the 1970s and 1980s essentially does this, with the addition that the numbering reflects argument behavior, e.g., 2s or direct objects are the NPs that are fronted when a verb is passivized.
 - PropBank takes a similar approach, numbering the roles ARG0, ARG1, ARG2, ARG3, (rarely other ARG#), plus ARGM (for adjuncts and modals)
 - Solution 2: Assume that roles are specific to frames, situations covered by a set of predicates including nouns, verbs and adjectives (FrameNet)
 - The **Contingency** frame: dependence (N), dependent (A), depend (V), determine (V), factor (N), function (N), hang on (V), hinge (V), independence (N), independent (A), ...
 - *Our success/**OUTCOME** may depend on available resources/**DETERMINANT***



Selection Restrictions

- If PRED selects ARG, PRED imposes some semantic restrictions on ARG
- For example, *ate* requires that its AGENT be alive and its PATIENT be a concrete object
- Imperfect relation between semantic roles and selection
 - One reason why semantic roles can proliferate as per previous slide
- A common computational approximation of selection restrictions:
 - Automatically acquired statistics about pairs of predicates and arguments
 - VERB SUBJECT: {[eat,John],[eat,Mary],[eat,lion],[eat,monster]...}
 - VERB OBJECT: {[eat,sandwich],[eat,fish],[eat,lasagna],[eat,clam]...}
 - Statistics-based weights, generalized using WordNet, etc.
- Statistical co-occurrence of heads of arguments can be used for:
 - Improving parsing: *The article about the wine [that we drank]*
 - *[[The article about the wine][that we drank]]* vs. *[the wine that we drank]*
 - Sense Disambiguation:
 - Given Pred selects Arg, favor pairs of senses that co-occur
 - *The bank lent me money*
 - » *lean* = borrow, not stand at an angle
 - » *bank* = organization, not side of river



PropBank

- PropBank consists of Lexicon and Annotation
- Lexicon
 - Each verb is associated with 1 or more rolesets (~senses)
 - Each roleset lists the **core** arguments of the verb
 - These are numbered: ARG0, ARG1, ARG2, ...
 - Additional information intended for annotation purposes
 - But possibly usable for other purposes as well (see later slide)
- Annotation: points to annotated PTB constituents
 - Verb
 - Core arguments: subset of {ARG0, ..., ARG9}
 - Non-core ARGM-XXX arguments:
 - -TMP, -LOC, -MNR, -CAU, -MOD, ...



PropBank Frame: <predicate lemma="purr">

- <roleset id="purr.01" name="emit purring noise" vncls="-">
 <roles> <role descr="cat" n="0"/ </roles>
 <example name="ewwwwwww">
 <text> John's bed purrs like a kitten </text>
 <arg n="0">John's bed</arg>
 <rel>purrs</rel>
 <arg f="MNR" n="m">like a kitten</arg> </example> </roleset>
- <roleset id="purr.02" name="speak in a purring fashion" vncls="37.3">
 <roles> <role descr="speaker" n="0"><vnrole vncls="37.3" vntheta="Agent"/></role>
 <role descr="utterance" n="1"><vnrole vncls="37.3" vntheta="Topic"/></role>
 <role descr="hearer" n="2"><vnrole vncls="37.3" vntheta="Recipient"/></role> </roles>
 <example name="automatically generated">
 <text> `` [It 's like an oasis in this room]-1 , " Ms. Foster purrs *trace*-1 .</text>
 <arg n="0">Ms. Foster</arg>
 <rel>purrs</rel>
 <arg n="1">*trace*</arg> </example></roleset>



Sample PropBank Annotation

- wsj/00/wsj_0041.mrg 57 13 gold purr.02 vn-3a 10:1-ARG0 13:0-rel 16:3-ARG1
 - File, tree num (0,...,N), token num (0,...,M), annotator, lemma.roleset, inflection, FV_1 , FV_2 , ... FV_N
 - FV = Token:Levels_above_the_leaf-ROLE (rel = predicate)
 - 10:1 = Phrase 1 level above the eleventh token
 - 10 = the, 10:1 = the female voice
 - ARG0 – the role described in the frame as SPEAKER

(S Pictures of rusted oil drums swim into focus and ,

(S (NP-SBJ (DT the) (JJ female) (NN voice))

(VP (VBZ purrs) (, ,) (` ``)

(FRAG

(NP (NP (DT That) (JJ hazardous) (NN waste))

(PP-LOC (IN on)

(NP (PRP\$ his)

(PRN (-LRB- -LCB-)

(NP (NNP Mr.) (NNP Courter) (POS 's))

(-RRB- -RCB-))

(NN property)))) ...)



Factors for PropBank SRL

- Identifying potential arguments for a verb instance
 - Depends on paths in parse trees
 - SRL without parsing achieves worse results
- Identifying the roleset for a given verb instance
 - Like sense disambiguation
 - Related to set of potential arguments
 - Related to set of reasonable labels
 - Some long-distance (gap filling) relations
- Labeling the arguments
 - Related to roleset (note inter-dependency)
 - Verb head/Noun head selection predicts choice
 - Grammatical position predicts choice
 - A role can be assigned to only one constituent, i.e., only one theme, agent, etc. (virtually all linguistic theories assume this)
 - Exceptions only apparent: *John ate rice and Mary pasta* (2 events)



Current PropBank SRL Systems

- Machine Learning: maximum entropy, etc.
- Features
 - Paths from arguments to verbs in parsing tree
 - *NP ↑S ↓VP ↓VBD*
 - This is the path from a subject up to the S node and down to the past tense head verb
 - Verb: POS, word, voice, subcategorization
 - Head of argument: POS, word, ...
 - Relative order between verb and argument
- Single classifiers vs. Breaking down problem into several classifiers (+/- argument, +/- core, ARG0 vs ARG1 ... vs. ARGN)



Further Reading on the PropBank SRL task

- Gildea and Jurafsky 2002
 - <http://www.cs.rochester.edu/~gildea/gildea-cl02.pdf>
- CONLL 2004 and 2005 – 2 shared tasks
 - <http://www.lsi.upc.edu/~srlconll/st04/st04.html>
 - <http://www.lsi.upc.edu/~srlconll/st05/st05.html>
- For Chinese (Xue and Palmer 2005)
 - <http://verbs.colorado.edu/~xuen/publications/ijcai05.pdf>



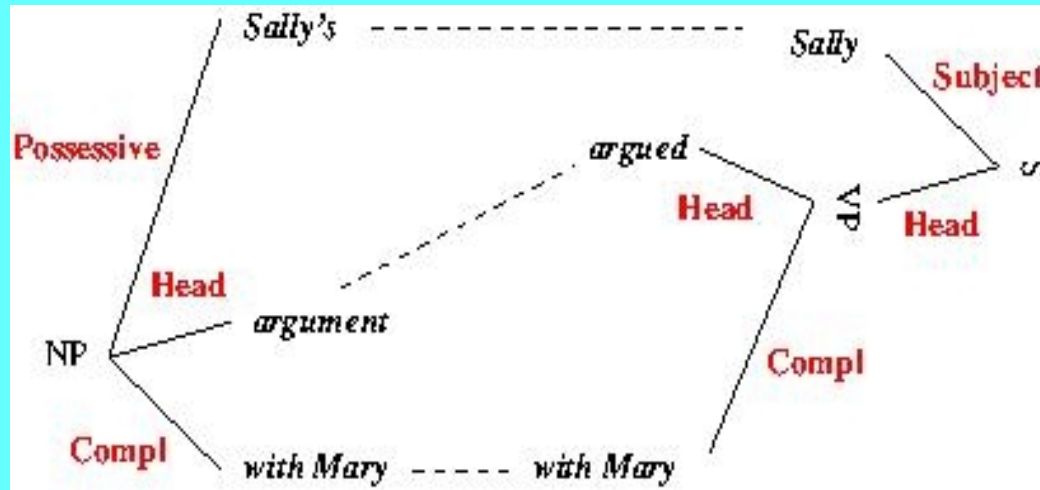
Arguments of NonVerbs

- A verb-centric bias
 - Most work on argument structure has always been about verbs
 - Sentences (and therefore verbs) are arguably more important than other phrasal units
 - Thematic Roles (Gruber and subsequent work) are biased towards sentential thematic role scenarios
- However, nouns and adjectives do take arguments
- Noun/Adj Arguments vary their positions in syntactically and lexically predicable ways
- Nouns are more frequent than verbs
 - Nouns can have similar arguments to related verbs (see next slide)
- Adjectives are less frequent (and vary less syntactically)
 - There has been little interest in funding an adjective argument task



Nomlex

- Dictionary mapping between NP headed by a noun N with Sentence with main verb V, if N and V are lexically related



- Macleod, Meyers, and others: <http://nlp.cs.nyu.edu/nomlex/index.html>
- NP positions: DET-POS (possessive), N-N-MOD (prenominal noun), PPs (with prepositions), etc.
- S positions: SBJ, OBJ, IND-OBJ, PPs, etc.
- For each subcategorization (in COMLEX entry for verb), all possible noun ↔ verb mappings are provided



Intro to NomBank

- <http://nlp.cs.nyu.edu/meyers/NomBank.html>
- Framework of NomBank
 - Modified version of PropBank framework to fit nominalizations
 - Uses frames from PropBank when possible for nouns related to verbs
- Expanded version of NOMLEX created semi-automatically to help preprocess data
- Creates classes of argument taking nouns which do not correspond to any verbs – the old set of semantic roles are not sufficient
- Deal with noun-specific phenomena: support constructions, transparent nouns and others

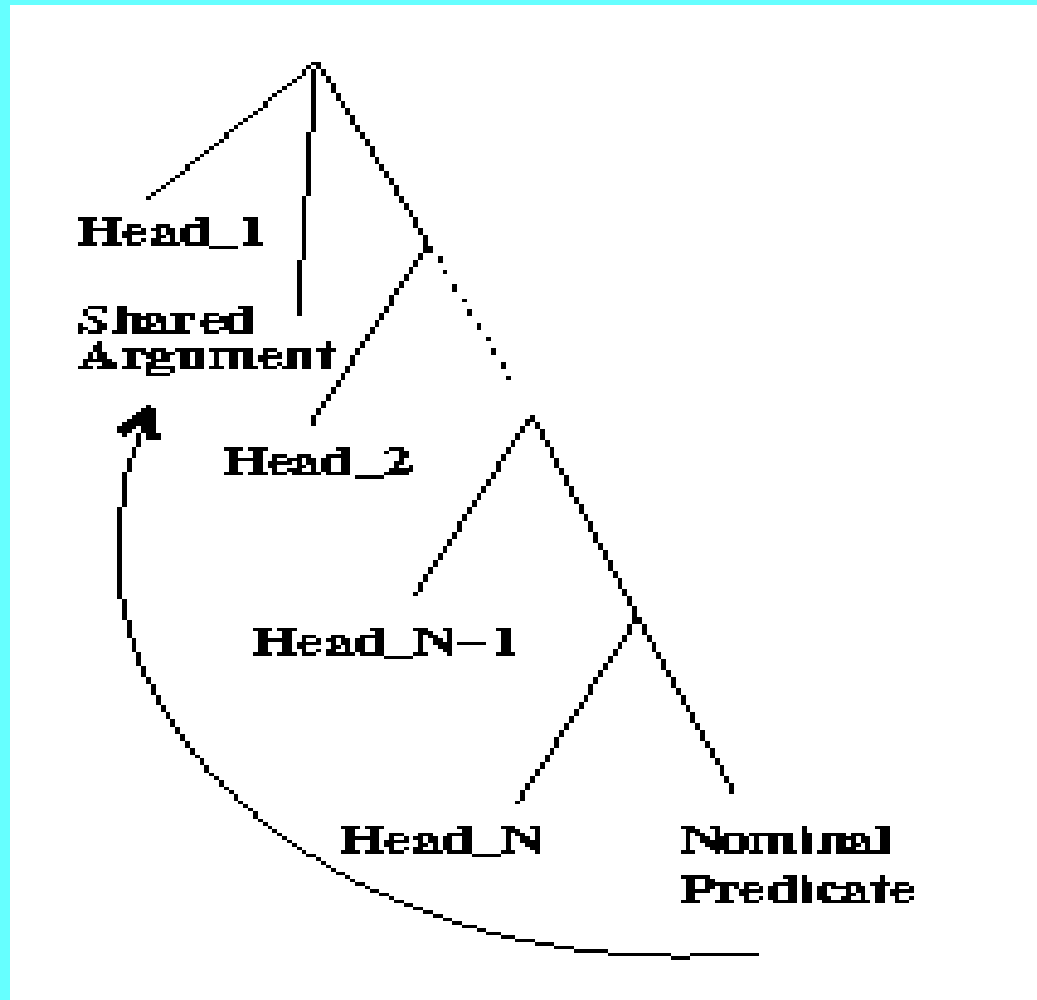


Noun Argument Structure Phenomena

- Noun Arguments ↔ Verb Arguments
 - *The aliens*/ARG0 *aided* *the humans*/ARG2 *with that project*/ARG1
 - *The aliens'* *aid* *for the humans on that project*
- Support
 - *The humans* *needed* *the aliens'* *aid* *on that project*
- Support Chains (typically support + transparent words)
 - *The humans* *needed* *a great deal of aid* *on that project*
- Adverbial Constructions
 - *With the aid of the aliens,* *the humans* *completed the project*
- Argument Nominalizations
 - *An admirer*/ARG0+REL *of benevolent aliens*/ARG1



Support Chain



Example Non-Verb-like Classes of Nouns

- *A period of industrial consolidation*/ARG1
 - [ENVIRONMENT]
- *Everyone*/ARG0 's *right to talk*/ARG1
 - [ABILITY]
- *Congress*/ARG0's *idea of reform*/ARG1
 - [WORK-OF-ART]
- *The topic of discussion*/ARG1
 - [CRISS-CROSS]
- *His*/ARG2 *math*/ARG1 *professor*/ARG0+REL
 - [RELATIONAL]
- *A set of tasks*/ARG1
 - [PARTITIVE]
- *The company*/ARG1 's *value of \$3 Billion*/ARG2
 - [ATTRIBUTE]



NomBank Frame

```
<roleset id="aid.01" name="help, aid" source="verb-aid.01">
<roles> <role descr="helper" n="0"></role>
        <role descr="project" n="1"></role>
        <role descr="beneficiary" n="2"></role></roles>
<example name="autogen1">
  <text> greater economic aid and technological know-how to
        flow from West to East </text>
  <arg f="EXT" n="M">greater</arg>
  <arg n="1">economic</arg>
  <rel>aid</rel>
  <arg n="Support">flow</arg>
  <arg n="0">from West</arg>
  <arg n="2">to East</arg> </example> </roleset>
```



NomBank Annotation

- wsj/01/wsj_0101.mrg 0 18 aid 01 17:0-ARG1 18:0-rel 19:1-ARG2
 - File, tree num (0,...,N), token num (0,...,M), lemma, roleset, FV_1 , FV_2 , ... FV_N
 - FV = Token:Levels_above_the_leaf-ROLE (rel = predicate)
 - (S ...A House-Senate conference approved major portions of a package for ...
(NP (NP
(QP (JJR more) (IN than) (\$ \$) (CD 500) (CD million))
(-NONE- *U*))
(PP (IN in)
(NP (NP (JJ economic) (NN aid))
(PP (IN for) (NP (NNP Poland)))))) ...)



NomBank vs. PropBank Annotation

- Completeness
 - PropBank: All 113,000 instances of main verbs and all their arguments were annotated (initially, not *be*).
 - NomBank: 114,500 out of 237,000 instances of common nouns were marked.
 - Only noun instances with arguments were marked.
 - Arguments included: ARG0, ..., ARGN and those ARGM classes that were found in PropBank. Other modifiers were not marked
- Argument Inventory
 - Both: ARG1, ... , ARGN, ARGM-XXX for several semantic classes
 - NomBank Only: Support verbs and Support Chains
- NomBank Only: A word can be its own argument
 - Argument nominalizations (so these can be mapped to verb frames)
- Exceptions to Rolesets = Senses
 - NomBank (rare, but happens): help, aid, ...
 - Argument and Verbal nominalizations use same roleset



More Info on NomBank

- All NYU's information including NomBank and related dictionaries, manuals and papers
 - <http://nlp.cs.nyu.edu/meyers/NomBank.html>
- Some NomBank SRL taggers
 - <http://www.comp.nus.edu.sg/~nght/pubs/emnlp06.pdf>
 - <http://aclweb.org/anthology/P/P10/P10-1160.pdf>
- CONLL 2008 and 2009
 - Shared Tasks incorporating multi-level parsing
 - The English (2008) task combines surface level dependency parsing with a logical level that combines PropBank/NomBank
 - 2009 is a multi-lingual task and resources vary among languages
 - Websites:
 - <http://barcelona.research.yahoo.net/dokuwiki/doku.php?id=conll2008:start>
 - <http://ufal.mff.cuni.cz/conll2009-st/>



Penn Discourse Treebank

- Captures relations between sentences in a PropBank-like framework
- Predicates: coordinate conjunctions (*and*, *but*, *or*), subordinate conjunctions (*if*, *while*, *when*, *since*), adverbs (*however*, *therefore*, *so*, *previously*), multi-word expressions in these classes (*as a result*, *as though*, *in contrast*)
- Two consecutive sentences can be in these same sort of discourse relations without having an explicit connective. These are annotated as if there is a virtual or implicit connective.
- Arguments:
 - ARG2 labels the phrase in the same clause as the connective
 - ARG1 labels the other phrase
- Examples:
 - *After the glass fell*/ARG2 , *it broke*/ARG1
 - *The glass broke*/ARG1 , *after it fell*/ARG2
 - *The glass broke.*/ARG1 *It had previously fallen*/ARG2
 - *The glass broke.*/ARG1 *It had fallen*/ARG2



More on the Penn Discourse Treebank

- Attribution phrases often must be skipped over
 - *Smoking makes you ill.*/ARG1 *Doctors say that, therefore you shouldn't smoke*/ARG2 .
- I am not aware of any statistically based PDTB labler
- My system (GLARF) automatically tags for explicit labels only and using simple manual rules – I have not done extensive evaluation



Some Concluding Notes on SRL

- Semantic role labeling
 - relates words and constituents semantically
 - ignores syntactic variation
- Most systems only handle verbs
- Some separate systems handle nouns
- CONLL 2008 and 2009 systems handle both
- Why not handle all predicates and arguments in one representation?
- What is the advantage of breaking these relations down one predicate type at a time?



Readings

- J & M: Chapter 19 and Section 20.9
- WordNet
 - Read the first 2 papers found here:
 - <http://wordnetcode.princeton.edu/5papers.pdf>
 - Read NLTK section 2.5 and try the NLTK WordNet module
- Optional Readings Throughout Slides

