Thesaurus-Based Similarity

Ling571
Deep Processing Techniques for NLP
February 22, 2017

Roadmap

- Lexical Semantics
  - Thesaurus-based Word Sense Disambiguation
    - Taxonomy-based similarity measures
    - Disambiguation strategies
  - Semantics summary
- Semantic Role Labeling
  - Task
  - Resources: PropBank, FrameNet
  - SRL systems

Previously

- Features for WSD:
  - Collocations, context, POS, syntactic relations
  - Can be exploited in classifiers
- Distributional semantics:
  - Vector representations of word "contexts"
    - Variable-sized windows
    - Dependency-relations
  - Similarity measures
- But, no prior knowledge of senses, sense relations

WordNet Taxonomy

- Most widely used English sense resource
- Manually constructed lexical database
  - 3 Tree-structured hierarchies
    - Nouns (117K) , verbs (11K), adjective+adverb (27K)
    - Entries: synonym set, gloss, example use
- Relations between entries:
  - Synonymy: in synset
  - Hypo(per)nym: Isa tree

WordNet

The noun "bass" has 8 senses in WordNet.

1. bass[1] - (the lowest part of the musical range)
2. bass[2], bass part[1] - (the lowest part in polyphonic music)
3. bass[3], basso - (an adult male singer with the lowest voice)
4. sea bass[1], bass[4] - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass[1], bass[5] - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6], bass voice[1], basso[2] - (the lowest adult male singing voice)
7. bass[7] - (the member with the lowest range of a family of musical instruments)
8. bass[8] - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

The adjective bass' has 1 sense in WordNet.

1. bass[1], deep[6] - (having or denoting a low vocal or instrumental range)
   "a deep voice"; 'a bass voice is lower than a baritone voice'; a bass clarinet

Noun WordNet Relations

Relation	Also Called	Definition	Exa1nple
Hypernym	Superordinate	From concepts to superordinates	*breakfast* [1] -----+ *rnea/*[1]
Hyponym	Subordinate	From concepts to subtypes	*rneaf*1 -----+ *lunch*[1]
Instance Hypernym	Instance	Fro111 instances to their concepts	*Austen*[1] -----+ *author*[1]
Instance Hyponym	Has-Instance	From concepts to concept instances	*cornposer*[1] -----+ *Bach*[1]
Member Meronym	Has-Member	From groups to their me1nbers	*facu!ty*[2] -----+ *projessor*[1]
Member Holonym	Member-Of	From members to their groups	*copilot*[1] -----+ *crew*[1]
Part Meronym	Has-Part	Fron1 wholes to parts	*table*[2] -----+ *leg*[3]
Part Holonym	Part-Of	From parts to wholes	*course*[7] ----+ *rneal1*
Substance Meronym		From substances to their subparts	*water*[1] -----+ *oxygen*[1]
Substance Holonym		From parts of substances to wholes	*gz˙n*[1] -----+ *nwrtz˙nz˙*[1]
Antonym		Semantic opposition between lemmas	*leader* [1] ¢===? *follower*[1]
Derivationally Related Form		Lenunas w/sa1ne morphological root	*destruction*[1] ¢===? *destro,*[1]

WordNet Taxonomy

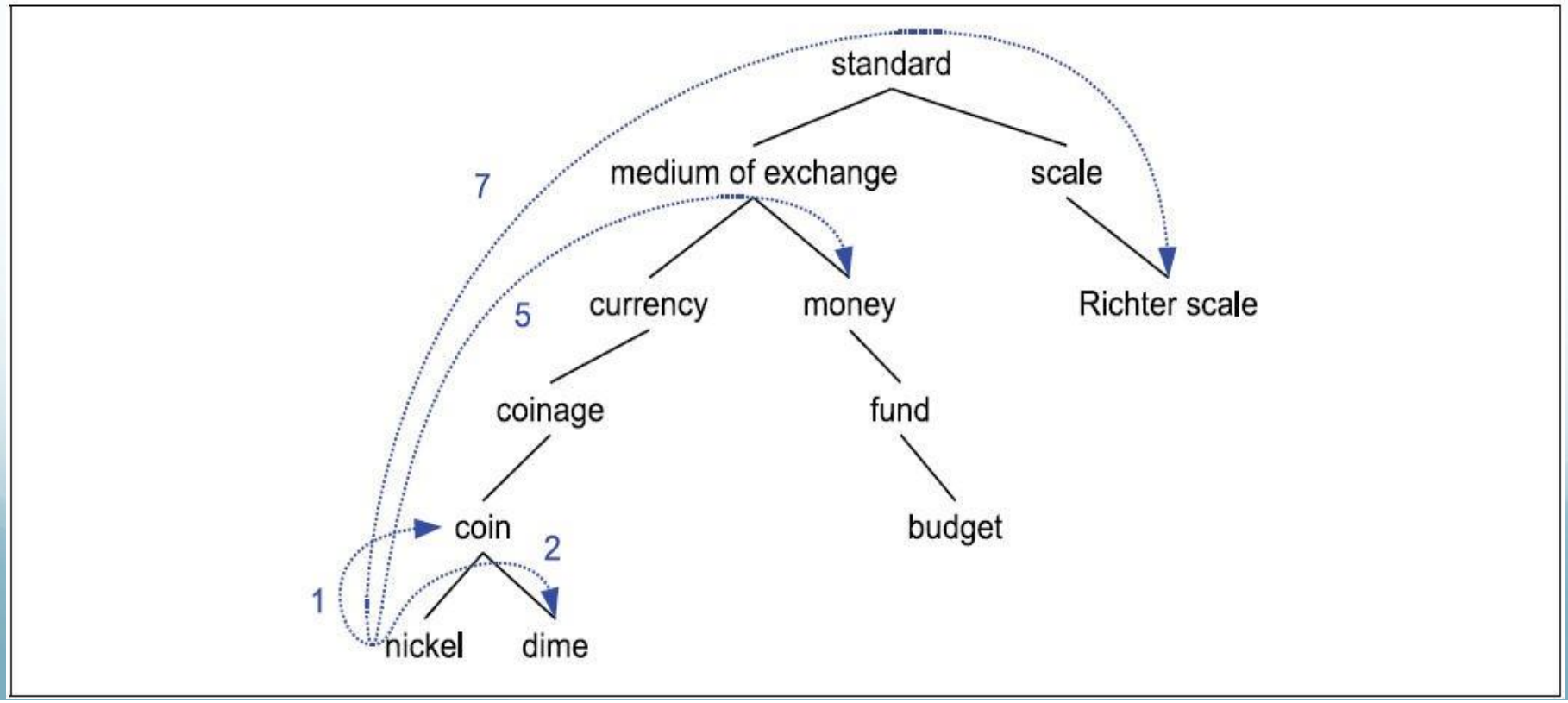
```
Sense 3
bass, bas.so
(an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, so:meo:ne..
               => or gan ism, being
                  => living thing, animate thing,
                     => whole unit
                        => object, physical object
                           => physical entity
                              => entity
               => causal agent, cause, causal agency
                  => physical entity
                     => entity
```

Thesaurus-based Techniques

- Key idea:
  - Shorter path length in thesaurus, smaller semantic dist.
    - Words similar to parents, siblings in tree
      - Further away, less similar
- Pathlength=# edges in shortest route in graph b/t nodes
  - Sim_{path}= -log pathlen(c_1 ,c_2) [Leacock & Chodorow]
- Problem 1:
  - Rarely know which sense, and thus which node
- Solution: assume most similar senses estimate
  - Wordsim(w_1,w_2) = max sim(c_1,c_2)

Path Length

- Path length problem:
  - Links in WordNet not uniform
    - Distance 5: Nickel->Money and Nickel->Standard



Information Content-Based Similarity Measures

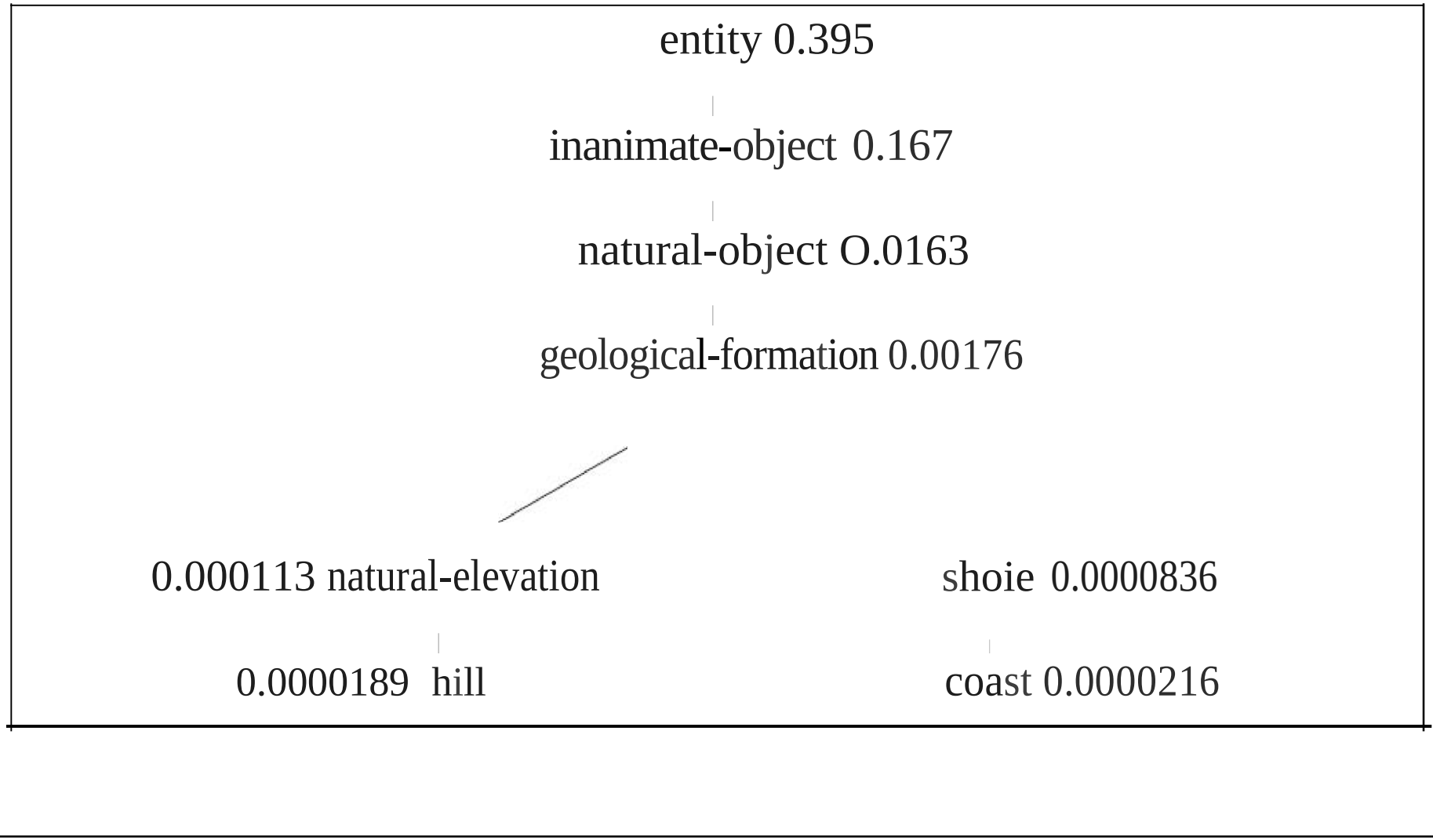
- Issues:
  - Word similarity vs sense similarity
    - Assume: sim(w1,w2) = $\max_{si:wi;sj:wj}$ (si,sj)
  - Path steps non-uniform
- Solution:
  - Add corpus information: information-content measure
    - P(c) : probability that a word is instance of concept c
      - Words(c) : words subsumed by concept c; N: words in corpus

$$P(c) = \frac{\sum_{w \in words(c)} count(w)}{N}$$

Information Content-Based Similarity Measures

- Information content of node:
  - $IC(c) = -\log P(c)$
- Least common subsumer (LCS):
  - Lowest node in hierarchy subsuming 2 nodes
- Similarity measure:
  - $sim_{RESNIK}(c_1,c_2) = -\log P(LCS(c_1,c_2))$

Concept Probability Example



Information Content-Based Similarity Measures

- Information content of node:
  - IC(c) = -log P(c)
- Least common subsumer (LCS):
  - Lowest node in hierarchy subsuming 2 nodes
- Similarity measure:
  - $sim_{RESNIK}(c_1,c_2)$ = - log $P(LCS(c_1,c_2))$
- Issue:
  - Not content, but difference between node & LCS

$$sim_{Lin}(c_1,c_2) = \frac{2 \times \log P(LCS(c_1,c_2))}{\log P(c_1) + \log P(c_2)}$$

Application to WSD

- Calculate Informativeness
  - For Each Node in WordNet:
    - Sum occurrences of concept and all children
    - Compute IC
- Disambiguate with WordNet
  - Assume set of words in context
    - E.g. {plants, animals, rainforest, species} from article
  - Find Most Informative Subsumer for each pair, I
    - Find LCS for each pair of senses, pick highest similarity
  - For each subsumed sense, Vote += I
  - Select Sense with Highest Vote

There are more kinds of plants and animals in the rainforests than anywhere else on Earth. Over half of the millions of known species of plants and animals live in the rainforest. Many are found nowhere else. There are even plants and animals in the rainforest that we have not yet discovered.

Biological Example

The Paulus company was founded in 1938. Since those days the product range has been the subject of constant expansions and is brought up continuously to correspond with the state of the art. We’re engineering, manufacturing and commissioning world-wide ready-to-run plants packed with our comprehensive know-how. Our Product Range includes pneumatic conveying systems for carbon, carbide, sand, lime and many others. We use reagent injection in molten metal for the…

Industrial Example

Label the First Use of “Plant”

Sense Labeling Under WordNet

- Use Local Content Words as Clusters
  - Biology: Plants, Animals, Rainforests, species…
  - Industry: Company, Products, Range, Systems…
- Find Common Ancestors in WordNet
  - Biology: Plants & Animals isa Living Thing
  - Industry: Product & Plant isa Artifact isa Entity
  - Use Most Informative
- Result: Correct Selection

Thesaurus Similarity Issues

- Coverage:
  - Few languages have large thesauri
  - Few languages have large sense tagged corpora
- Thesaurus design:
  - Works well for noun IS-A hierarchy
  - Verb hierarchy shallow, bushy, less informative

Semantic Role Labeling

Roadmap

- Semantic role labeling (SRL):
  - Motivation:
    - Between deep semantics and slot-filling
  - Thematic roles
  - Thematic role resources
    - PropBank, FrameNet
  - Automatic SRL approaches

Semantic Analysis

- Two extremes:
  - Full, deep compositional semantics
    - Creates full logical form
    - Links sentence meaning representation to logical world model representation
    - Powerful, expressive, AI-complete
  - Domain-specific slot-filling:
    - Common in dialog systems, IE tasks
    - Narrowly targeted to domain/task
    - Often pattern-matching
    - Low cost, but lacks generality, richness, etc

Semantic Role Labeling

- Typically want to know:
  - *Who* did *what* to *whom*, *where*, *when*, and *how*
- Intermediate level:
  - Shallower than full deep composition
  - Abstracts away (somewhat) from surface form
  - Captures general predicate-argument structure info
  - Balance generality and specificity

Example

- Yesterday Tom chased Jerry.
- Yesterday Jerry was chased by Tom.
- Tom chased Jerry yesterday.
- Jerry was chased yesterday by Tom.

- Semantic roles:
  - Chaser: Tom
  - ChasedThing: Jerry
  - TimeOfChasing: yesterday
- Same across all sentence forms

Full Event Semantics

- Neo-Davidsonian style:
  - exists e. Chasing(e) & Chaser(e,Tom) & ChasedThing(e,Jerry) & TimeOfChasing(e,Yesterday)
- Same across all examples
- Roles: Chaser, ChasedThing, TimeOfChasing
  - Specific to verb "chase"
  - Aka "Deep roles"

Issues

- Challenges:
  - How many roles for a language?
    - Arbitrarily many deep roles
      - Specific to each verb's event structure
  - How can we acquire these roles?
    - Manual construction?
    - Some progress on automatic learning
      - Still only successful on limited domains (ATIS, geography)
  - Can we capture generalities across verbs/events?
    - Not really, each event/role is specific
- Alternative: thematic roles

Thematic Roles

- Describe semantic roles of verbal arguments
  - Capture commonality across verbs
  - E.g. subject of break, open is AGENT
    - AGENT: volitional cause
    - THEME: things affected by action
  - Enables generalization over surface order of arguments
    - $\text{John}_{\text{AGENT}}$ broke the $\text{window}_{\text{THEME}}$
    - The $\text{rock}_{\text{INSTRUMENT}}$ broke the $\text{window}_{\text{THEME}}$
    - The $\text{window}_{\text{THEME}}$ was broken by $\text{John}_{\text{AGENT}}$

Thematic Roles

- Thematic grid, θ-grid, case frame
  - Set of thematic role arguments of verb
    - E.g. Subject: AGENT; Object: THEME, or
    - Subject: INSTR; Object: THEME

- Verb/Diathesis Alternations
  - Verbs allow different surface realizations of roles
    - $\text{Doris}_{\text{AGENT}}$ gave the $\text{book}_{\text{THEME}}$ to $\text{Cary}_{\text{GOAL}}$
    - $\text{Doris}_{\text{AGENT}}$ gave $\text{Cary}_{\text{GOAL}}$ the $\text{book}_{\text{THEME}}$
  - Group verbs into classes based on shared patterns

Canonical Roles

Thematic Role	Example
AGENT	*The waiter* spilled the soup.
EXPERIENCER	*John* has a headache.
FORCE	*The wind* blo,vs debris from the 1nall into our yards.
THEME	Only after Benjamin Franklin broke *the ice*...
RESULT	The French goverrunent has built a *regulation-size baseball dianlond*...
CONTENT	Mona asked *'Yournet i\1ary Ann at a super1narket?'*
INSTRU fENT	He turned to poaching catfish, stunning them *with a shocking device* ...
BENEF1CIARY	Whenever Ann Callahan makes hotel reservations */or her boss*...
SOURCE	Ifle in *frorrt Boston.*
GOAL	I drove *to Portland.*

Thematic Role Issues

- Hard to produce
  - Standard set of roles
    - Fragmentation: Often need to make more specific
      - E,g, INSTRUMENTS can be subject or not
  - Standard definition of roles
    - Most AGENTs: animate, volitional, sentient, causal
    - But not all….
- Strategies:
  - Generalized semantic roles: PROTO-AGENT/PROTO-PATIENT
    - Defined heuristically: PropBank
  - Define roles specific to verbs/nouns: FrameNet

PropBank

- Sentences annotated with semantic roles
  - Penn and Chinese Treebank
  - Roles specific to verb sense
    - Numbered: Arg0, Arg1, Arg2,…
      - Arg0: PROTO-AGENT; Arg1: PROTO-PATIENT, etc
      - >1: Verb-specific
  - E.g. agree.01
    - Arg0: Agreer
    - Arg1: Proposition
    - Arg2: Other entity agreeing
    - Ex1: [$_{Arg0}$The group] agreed [$_{Arg1}$it wouldn't make an offer]

Propbank

- Resources:
  - Annotated sentences
    - Started w/Penn Treebank
    - Now: Google answerbank, SMS, webtext, etc
      - Also English and Arabic
  - Framesets:
    - Per-sense inventories of roles, examples
    - Span verbs, adjectives, nouns (e.g. event nouns)
- http://verbs.colorado.edu/propbank
- Recent status:
  - 5940 verbs w/ 8121 framesets;
  - 1880 adjectives w/2210 framesets

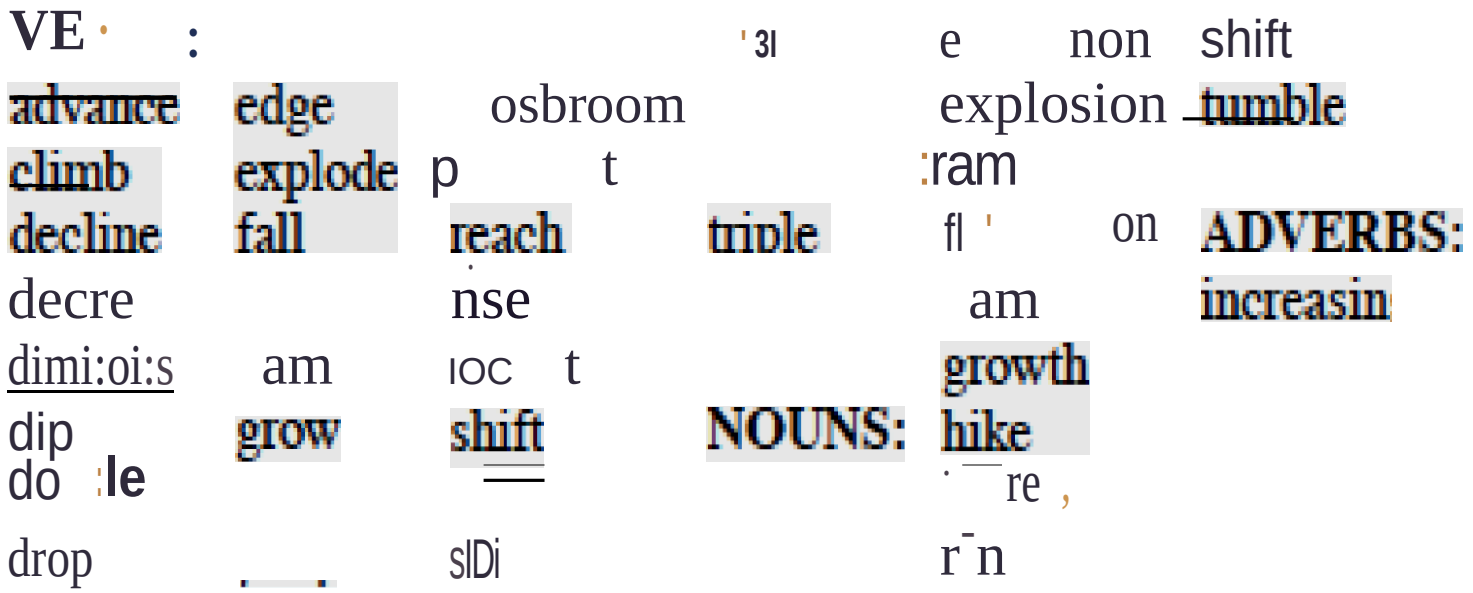
FrameNet (Fillmore et al)

- Key insight:
  - Commonalities not just across diff't sentences w/*same* verb but across *different* verbs (and nouns and adjs)
- PropBank
  - [$_{Arg0}$Big Fruit Co.] increased [$_{Arg1}$ the price of bananas].
  - [$_{Arg1}$The price of bananas] was increased by [$_{Arg0}$ BFCo].
  - [$_{Arg1}$The price of bananas] increased [$_{Arg2}$ 5%].
- FrameNet
  - [$_{ATTRIBUTE}$The price] of [$_{ITEM}$bananas] increased [$_{DIFF}$5%].
  - [$_{ATTRIBUTE}$The price] of [$_{ITEM}$bananas] rose [$_{DIFF}$5%].
  - There has been a [$_{DIFF}$5%] rise in [$_{ATTRIBUTE}$ the price] of [$_{ITEM}$ bananas].

FrameNet

- Semantic roles specific to Frame
  - Frame: script-like structure, roles (frame elements)
  - E.g. change_position_on_scale: increase, rise
    - Attribute, Initial_value, Final_value
  - Core, non-core roles
  - Relationships b/t frames, frame elements
    - Add causative: cause_change_position_on_scale

Change of position on scale



Core Roles

ATTRIBUTE	The ATTRIBUTE is a scalar property that the ITEM possesses.
DIFFERENCE	The distance by which an ITEM changes its position on the scale.
FINAL_STATE	A description that presents the ITEM's state after the change in the ATTRIBUTE's value as an independent predication.
FINAL_VALUE	The position on the scale \\'here the ITEM ends up.
INITIAL_STATE	A description that presents the ITEM's state before the change in the ATTRIBUTE's value as an independent predication.
INITIAL_ VALUE	The initial position on the scale fro1n which the ITEM 1noves a\\1ay.
ITEM	The entity that has a position on the scale.
VALUE_RANGE	A portion of the scale, typically identified by its end points, along \vhich the values of the ATTRIBUTE fluctuate.

Some Non-Core Roles

DURATION	The length of tin1e over ,;vhich the change takes place.
SPEED	The rate of change of the VALUE.
GROUP	The GROUP in \vhich an ITEM changes the value of an ATTRIBUTE in a specified way.

FrameNet

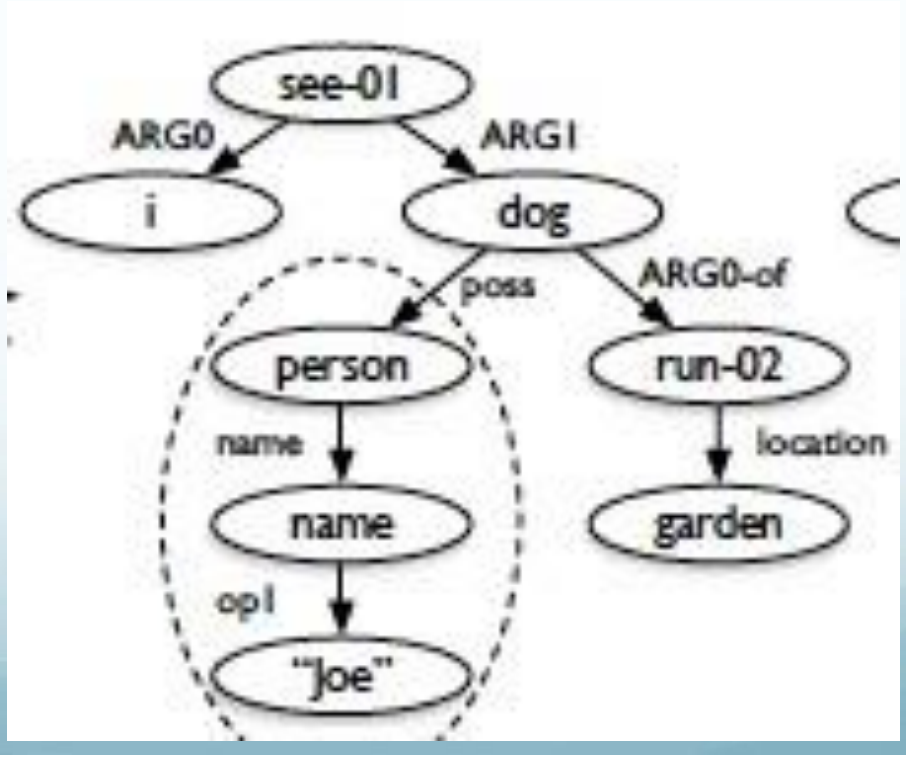
- Current status:
  - 1222 frames
  - ~13500 lexical units (mostly verbs, nouns)
  - Annotations over:
    - Newswire (WSJ, AQUAINT)
    - American National Corpus
- Under active development
- Still only ~6K verbs, limited coverage

AMR

- "Abstract Meaning Representation"
  - Sentence-level semantic representation
  - Nodes: Concepts:
    - English words, PropBank predicates, or keywords ('person')
  - Edges: Relations:
    - PropBank thematic roles (ARG0-ARG5)
    - Others including 'location', 'name', 'time', etc…
    - ~100 in total

AMR 2

- AMR Bank: (now) ~40K annotated sentences
- JAMR parser: 63% F-measure (2015)
  - Alignments b/t word spans & graph fragments
- Example: "I saw Joe's dog, which was running in the garden."



Liu et al, 2015.