Lexical Semantics

Ling571
Deep Processing Techniques for NLP
February 23, 2015

What is a plant?

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 **plant**s and animals in the rainforest that we have not yet discovered.

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.

Lexical Semantics

- So far, word meanings discrete
  - Constants, predicates, functions

Lexical Semantics

- So far, word meanings discrete
  - Constants, predicates, functions
- Focus on word meanings:
  - Relations of meaning among words
    - Similarities & differences of meaning in sim context

Lexical Semantics

- So far, word meanings discrete
  - Constants, predicates, functions
- Focus on word meanings:
  - Relations of meaning among words
    - Similarities & differences of meaning in sim context
  - Internal meaning structure of words
    - Basic internal units combine for meaning

Terminology

- **Lexeme**:
  - Form: Orthographic/phonological + meaning

Terminology

- **Lexeme**:
  - Form: Orthographic/phonological + meaning
  - Represented by lemma
    - **Lemma**: citation form; infinitive in inflection
      - Sing: sing, sings, sang, sung,…

Terminology

- **Lexeme**:
  - Form: Orthographic/phonological + meaning
  - Represented by lemma
    - **Lemma**: citation form; infinitive in inflection
      - Sing: sing, sings, sang, sung,…
- **Lexicon**: finite list of lexemes

Sources of Confusion

- Homonymy:
  - Words have same form but different meanings
    - Generally same POS, but unrelated meaning

Sources of Confusion

- Homonymy:
  - Words have same form but different meanings
    - Generally same POS, but unrelated meaning
    - E.g. bank (side of river) vs bank (financial institution)
      - bank^1 vs bank^2

Sources of Confusion

- Homonymy:
  - Words have same form but different meanings
    - Generally same POS, but unrelated meaning
    - E.g. bank (side of river) vs bank (financial institution)
      - $bank^1$ vs $bank^2$
    - Homophones: same phonology, diff't orthographic form
      - E.g. two, to, too

Sources of Confusion

- Homonymy:
  - Words have same form but different meanings
    - Generally same POS, but unrelated meaning
    - E.g. bank (side of river) vs bank (financial institution)
      - $bank^1$ vs $bank^2$
    - Homophones: same phonology, diff't orthographic form
      - E.g. two, to, too
    - Homographs: Same orthography, diff't phonology
- Why?

Sources of Confusion

- Homonymy:
  - Words have same form but different meanings
    - Generally same POS, but unrelated meaning
    - E.g. bank (side of river) vs bank (financial institution)
      - $bank^1$ vs $bank^2$
    - Homophones: same phonology, diff't orthographic form
      - E.g. two, to, too
    - Homographs: Same orthography, diff't phonology
- Why?
  - Problem for applications: TTS, ASR transcription, IR

Sources of Confusion II

- Polysemy
  - Multiple RELATED senses
    - E.g. bank: money, organ, blood,...

Sources of Confusion II

- Polysemy
  - Multiple RELATED senses
    - E.g. bank: money, organ, blood,…
  - Big issue in lexicography
    - # of senses, relations among senses, differentiation
    - E.g. serve breakfast, serve Philadelphia, serve time

Relations between Senses

- Synonymy:
  - (near) identical meaning

Relations between Senses

- Synonymy:
  - (near) identical meaning
  - Substitutability
    - Maintains propositional meaning
  - Issues:

Relations between Senses

- Synonymy:
  - (near) identical meaning
  - Substitutability
    - Maintains propositional meaning
  - Issues:
    - Polysemy – same as some sense

Relations between Senses

- Synonymy:
  - (near) identical meaning
  - Substitutability
    - Maintains propositional meaning
  - Issues:
    - Polysemy – same as some sense
    - Shades of meaning – other associations:
      - Price/fare; big/large; water H_2O

Relations between Senses

- Synonymy:
  - (near) identical meaning
  - Substitutability
    - Maintains propositional meaning
  - Issues:
    - Polysemy – same as some sense
    - Shades of meaning – other associations:
      - Price/fare; big/large; water H_2O
    - Collocational constraints: e.g. babbling brook

Relations between Senses

- Synonymy:
  - (near) identical meaning
  - Substitutability
    - Maintains propositional meaning
  - Issues:
    - Polysemy – same as some sense
    - Shades of meaning – other associations:
      - Price/fare; big/large; water H_2O
    - Collocational constraints: e.g. babbling brook
    - Register:
      - social factors: e.g. politeness, formality

Relations between Senses

- Antonyms:
  - Opposition
    - Typically ends of a scale
      - Fast/slow; big/little

Relations between Senses

- Antonyms:
  - Opposition
    - Typically ends of a scale
      - Fast/slow; big/little
    - Can be hard to distinguish automatically from syns

Relations between Senses

- Antonyms:
  - Opposition
    - Typically ends of a scale
      - Fast/slow; big/little
    - Can be hard to distinguish automatically from syns
- Hyponomy:
  - Isa relations:
    - More General (hypernym) vs more specific (hyponym)
      - E.g. dog/golden retriever; fruit/mango;

Relations between Senses

- Antonyms:
  - Opposition
    - Typically ends of a scale
      - Fast/slow; big/little
    - Can be hard to distinguish automatically from syns
- Hyponomy:
  - Isa relations:
    - More General (hypernym) vs more specific (hyponym)
      - E.g. dog/golden retriever; fruit/mango;
  - Organize as ontology/taxonomy

WordNet Taxonomy

- Most widely used English sense resource
- Manually constructed lexical database

WordNet Taxonomy

- Most widely used English sense resource
- Manually constructed lexical database
  - 3 Tree-structured hierarchies
    - Nouns (117K) , verbs (11K), adjective+adverb (27K)

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

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. $bass^3$, $basso^1$ - (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	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*^{1} → *meal*^{1}
Hyponym	Subordinate	From concepts to subtypes	*meal*^{1} → *lunch*^{1}
Instance Hypernym	Instance	From instances to their concepts	*Austen*^{1} → *author*^{1}
Instance Hyponym	Has-Instance	From concepts to concept instances	*composer*^{1} → *Bach*^{1}
Member Meronym	Has-Member	From groups to their members	*faculty*^{2} → *professor*^{1}
Member Holonym	Member-Of	From members to their groups	*copilot*^{1} → *crew*^{1}
Part Meronym	Has-Part	From wholes to parts	*table*^{2} → *leg*^{3}
Part Holonym	Part-Of	From parts to wholes	*course*^{7} → *meal*^{1}
Substance Meronym		From substances to their subparts	*water*^{1} → *oxygen*^{1}
Substance Holonym		From parts of substances to wholes	*gin*^{1} → *martini*^{1}
Antonym		Semantic opposition between lemmas	*leader*^{1} ⟺ *follower*^{1}
Derivationally Related Form		Lemmas w/same morphological root	*destruction*^{1} ⟺ *destroy*^{1}

WordNet Taxonomy

```
Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
               => causal agent, cause, causal agency
                  => physical entity
                     => entity
```

Word Sense Disambiguation

- WSD
  - Tasks, evaluation, features
  - Selectional Restriction-based Approaches
  - Robust Approaches
    - Dictionary-based Approaches
    - Distributional Approaches
    - Resource-based Approaches
- Summary
  - Strengths and Limitations

Word Sense Disambiguation

- Application of lexical semantics
- Goal: Given a word *in context*, identify the appropriate sense
  - E.g. plants and animals in the rainforest
- Crucial for real syntactic & semantic analysis

Word Sense Disambiguation

- Application of lexical semantics
- Goal: Given a word *in context*, identify the appropriate sense
  - E.g. <u>plants</u> and animals in the rainforest
- Crucial for real syntactic & semantic analysis
  - Correct sense can determine
    - .

Word Sense Disambiguation

- Application of lexical semantics
- Goal: Given a word *in context*, identify the appropriate sense
  - E.g. <u>plants</u> and animals in the rainforest
- Crucial for real syntactic & semantic analysis
  - Correct sense can determine
    - Available syntactic structure
    - Available thematic roles, correct meaning,..

Robust Disambiguation

- More to semantics than P-A structure
  - Select sense where predicates underconstrain

Robust Disambiguation

- More to semantics than P-A structure
  - Select sense where predicates underconstrain
- Learning approaches
  - Supervised, Bootstrapped, Unsupervised

Robust Disambiguation

- More to semantics than P-A structure
  - Select sense where predicates underconstrain
- Learning approaches
  - Supervised, Bootstrapped, Unsupervised
- Knowledge-based approaches
  - Dictionaries, Taxonomies
- Widen notion of context for sense selection

Robust Disambiguation

- More to semantics than P-A structure
  - Select sense where predicates underconstrain
- Learning approaches
  - Supervised, Bootstrapped, Unsupervised
- Knowledge-based approaches
    - Dictionaries, Taxonomies
- Widen notion of context for sense selection
  - Words within window (2,50,discourse)
  - Narrow cooccurrence - collocations

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"

Disambiguation Features

- Key: What are the features?

Disambiguation Features

- Key: What are the features?
  - Part of speech
    - Of word and neighbors

Disambiguation Features

- Key: What are the features?
  - Part of speech
    - Of word and neighbors
  - Morphologically simplified form

Disambiguation Features

- Key: What are the features?
  - Part of speech
    - Of word and neighbors
  - Morphologically simplified form
  - Words in neighborhood

Disambiguation Features

- Key: What are the features?
  - Part of speech
    - Of word and neighbors
  - Morphologically simplified form
  - Words in neighborhood
    - Question: How big a neighborhood?

Disambiguation Features

- Key: What are the features?
  - Part of speech
    - Of word and neighbors
  - Morphologically simplified form
  - Words in neighborhood
    - Question: How big a neighborhood?
      - Is there a single optimal size? Why?
  - ..

Disambiguation Features

- Key: What are the features?
  - Part of speech
    - Of word and neighbors
  - Morphologically simplified form
  - Words in neighborhood
    - Question: How big a neighborhood?
      - Is there a single optimal size? Why?
  - (Possibly shallow) Syntactic analysis
    - E.g. predicate-argument relations, modification, phrases
  - Collocation vs co-occurrence features

Disambiguation Features

- Key: What are the features?
  - Part of speech
    - Of word and neighbors
  - Morphologically simplified form
  - Words in neighborhood
    - Question: How big a neighborhood?
      - Is there a single optimal size? Why?
  - (Possibly shallow) Syntactic analysis
    - E.g. predicate-argument relations, modification, phrases
  - Collocation vs co-occurrence features
    - Collocation: words in specific relation: p-a, 1 word +/-

Disambiguation Features

- Key: What are the features?
  - Part of speech
    - Of word and neighbors
  - Morphologically simplified form
  - Words in neighborhood
    - Question: How big a neighborhood?
      - Is there a single optimal size? Why?
  - (Possibly shallow) Syntactic analysis
    - E.g. predicate-argument relations, modification, phrases
  - Collocation vs co-occurrence features
    - Collocation: words in specific relation: p-a, 1 word +/-
    - Co-occurrence: bag of words..

WSD Evaluation

- Ideally, end-to-end evaluation with WSD component
  - Demonstrate real impact of technique in system

WSD Evaluation

- Ideally, end-to-end evaluation with WSD component
  - Demonstrate real impact of technique in system
  - Difficult, expensive, still application specific

WSD Evaluation

- Ideally, end-to-end evaluation with WSD component
  - Demonstrate real impact of technique in system
  - Difficult, expensive, still application specific
- Typically, intrinsic, sense-based
  - Accuracy, precision, recall
  - SENSEVAL/SEMEVAL: all words, lexical sample
- Baseline:

WSD Evaluation

- Ideally, end-to-end evaluation with WSD component
  - Demonstrate real impact of technique in system
  - Difficult, expensive, still application specific
- Typically, intrinsic, sense-based
  - Accuracy, precision, recall
  - SENSEVAL/SEMEVAL: all words, lexical sample
- Baseline:
  - Most frequent sense, Lesk
- Topline:

WSD Evaluation

- Ideally, end-to-end evaluation with WSD component
  - Demonstrate real impact of technique in system
  - Difficult, expensive, still application specific
- Typically, intrinsic, sense-based
  - Accuracy, precision, recall
  - SENSEVAL/SEMEVAL: all words, lexical sample
- Baseline:
  - Most frequent sense, Lesk
- Topline:
  - Human inter-rater agreement: 75-80% fine; 90% coarse

Dictionary-Based Approach

- (Simplified) Lesk algorithm
  - “How to tell a pine cone from an ice cream cone”

Dictionary-Based Approach

- (Simplified) Lesk algorithm
  - “How to tell a pine cone from an ice cream cone”

- Compute ‘signature’ of word senses:
  - Words in gloss and examples in dictionary

Dictionary-Based Approach

- (Simplified) Lesk algorithm
  - "How to tell a pine cone from an ice cream cone"

- Compute 'signature' of word senses:
  - Words in gloss and examples in dictionary

- Compute context of word to disambiguate
  - Words in surrounding sentence(s)

Dictionary-Based Approach

- (Simplified) Lesk algorithm
  - "How to tell a pine cone from an ice cream cone"

- Compute 'signature' of word senses:
  - Words in gloss and examples in dictionary

- Compute context of word to disambiguate
  - Words in surrounding sentence(s)

- Compare overlap b/t signature and context

Dictionary-Based Approach

- (Simplified) Lesk algorithm
  - "How to tell a pine cone from an ice cream cone"
- Compute 'signature' of word senses:
  - Words in gloss and examples in dictionary
- Compute context of word to disambiguate
  - Words in surrounding sentence(s)
- Compare overlap b/t signature and context
- Select sense with highest (non-stopword) overlap

Applying Lesk

- *The bank can guarantee deposits will eventually cover future tuition costs because it invests in mortgage securities.*

$bank^1$	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
$bank^2$	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

Applying Lesk

- *The bank can guarantee deposits will eventually cover future tuition costs because it invests in mortgage securities.*

$bank^1$	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
$bank^2$	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

- $Bank^1$: 2

Applying Lesk

- *The bank can guarantee deposits will eventually cover future tuition costs because it invests in mortgage securities.*

bank^1	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
bank^2	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

- Bank^1 : 2
- Bank^2: 0

Improving Lesk

- Overlap score:
  - All words equally weighted (excluding stopwords)

Improving Lesk

- Overlap score:
  - All words equally weighted (excluding stopwords)

- Not all words equally informative

Improving Lesk

- Overlap score:
  - All words equally weighted (excluding stopwords)

- Not all words equally informative
  - Overlap with unusual/specific words – better
  - Overlap with common/non-specific words – less good

Improving Lesk

- Overlap score:
  - All words equally weighted (excluding stopwords)

- Not all words equally informative
  - Overlap with unusual/specific words – better
  - Overlap with common/non-specific words – less good
- Employ corpus weighting:
  - IDF: inverse document frequency
    - $Idf_i = \log(Ndoc/nd_i)$

Word Similarity

- Synonymy:

Word Similarity

- Synonymy:
  - True propositional substitutability is rare, slippery

Word Similarity

- Synonymy:
  - True propositional substitutability is rare, slippery
- Word similarity (semantic distance):
  - Looser notion, more flexible

Word Similarity

- Synonymy:
  - True propositional substitutability is rare, slippery
- Word similarity (semantic distance):
  - Looser notion, more flexible
  - Appropriate to applications:
    - IR, summarization, MT, essay scoring

Word Similarity

- Synonymy:
  - True propositional substitutability is rare, slippery
- Word similarity (semantic distance):
  - Looser notion, more flexible
  - Appropriate to applications:
    - IR, summarization, MT, essay scoring
      - Don't need binary +/- synonym decision

Word Similarity

- Synonymy:
  - True propositional substitutability is rare, slippery
- Word similarity (semantic distance):
  - Looser notion, more flexible
  - Appropriate to applications:
    - IR, summarization, MT, essay scoring
      - Don't need binary +/- synonym decision
      - Want terms/documents that have high similarity

Word Similarity

- Synonymy:
  - True propositional substitutability is rare, slippery
- Word similarity (semantic distance):
  - Looser notion, more flexible
  - Appropriate to applications:
    - IR, summarization, MT, essay scoring
      - Don't need binary +/- synonym decision
      - Want terms/documents that have high similarity
        - Differ from relatedness

Word Similarity

- Synonymy:
  - True propositional substitutability is rare, slippery
- Word similarity (semantic distance):
  - Looser notion, more flexible
  - Appropriate to applications:
    - IR, summarization, MT, essay scoring
      - Don't need binary +/- synonym decision
      - Want terms/documents that have high similarity
        - Differ from relatedness
- Approaches:
  - Thesaurus-based
  - Distributional

Distributional Similarity

- Unsupervised approach:
  - Clustering, WSD, automatic thesaurus enrichment

Some slides based on Eisenstein 2014

Distributional Similarity

- Unsupervised approach:
  - Clustering, WSD, automatic thesaurus enrichment
- Insight:
  - "You shall know a word by the company it keeps!"
    - (Firth, 1957)

Distributional Similarity

- Unsupervised approach:
  - Clustering, WSD, automatic thesaurus enrichment
- Insight:
  - “You shall know a word by the company it keeps!”
    - (Firth, 1957)
  - A bottle of *tezguino* is on the table.
  - Everybody likes *tezguino*.
  - *Tezguino* makes you drunk.
  - We make *tezguino* from corn.

Distributional Similarity

- Unsupervised approach:
  - Clustering, WSD, automatic thesaurus enrichment
- Insight:
  - "You shall know a word by the company it keeps!"
    - (Firth, 1957)
  - A bottle of *tezguino* is on the table.
  - Everybody likes *tezguino*.
  - *Tezguino* makes you drunk.
  - We make *tezguino* from corn.
- Tezguino: corn-based, alcoholic beverage

Local Context Clustering

- "Brown" (aka IBM) clustering (1992)
  - Generative model over adjacent words
  - Each w_i has class c_i
  - $\log P(W) = \Sigma_i \log P(w_i|c_i) + \log P(c_i|c_{i-1})$
    - (Familiar??)

Local Context Clustering

- "Brown" (aka IBM) clustering (1992)
  - Generative model over adjacent words
  - Each w_i has class c_i
  - $\log P(W) = \Sigma_i \log P(w_i|c_i) + \log P(c_i|c_{i-1})$
    - (Familiar??)
  - Greedy clustering
    - Start with each word in own cluster
    - Merge clusters based on log prob of text under model
      - Merge those which maximize P(W)

Clustering Impact

- Improves downstream tasks
  - Here Named Entity Recognition vs HMM (Miller et al '04)