CS 4650/7650, Lecture 3: Word-sense disambiguation

Jacob Eisenstein

August 27, 2013

Word senses

Remember these headlines?

- Iraqi head seeks arms
- Prostitutes appeal to Pope
- Drunk gets nine years in violin case

Word senses

Remember these headlines?

- Iraqi head seeks arms
- Prostitutes appeal to Pope
- Drunk gets nine years in violin case

They are ambiguous because words have multiple senses.

- head: BODY-PART,LEADER
- arms: BODY-PART,WEAPON

Word senses

Remember these headlines?

- Iraqi head seeks arms
- Prostitutes appeal to Pope
- Drunk gets nine years in violin case

They are ambiguous because words have multiple senses.

- head: BODY-PART,LEADER
- arms: BODY-PART,WEAPON

Can you see what is ambiguous about the other examples?

Word sense disambiguation

Word Sense Disambiguation (WSD) is the problem of identifying the intended sense of each word token.

- Part of a larger field of research called *lexical semantics*

Word sense disambiguation

Word Sense Disambiguation (WSD) is the problem of identifying the intended sense of each word token.

- Part of a larger field of research called *lexical semantics*
- Part-of-speech ambiguity (i'm heading out of town) is usually considered to be a different problem.

Word sense disambiguation

Word Sense Disambiguation (WSD) is the problem of identifying the intended sense of each word token.

- Part of a larger field of research called *lexical semantics*
- Part-of-speech ambiguity (i'm heading out of town) is usually considered to be a different problem.
- For WSD, you can think of words as including their POS tag (e.g., heading/V)

Word sense disambiguation

Word Sense Disambiguation (WSD) is the problem of identifying the intended sense of each word token.

- Part of a larger field of research called *lexical semantics*
- Part-of-speech ambiguity (i'm heading out of town) is usually considered to be a different problem.
- For WSD, you can think of words as including their POS tag (e.g., heading/V)
- Technically, we want to differentiate senses of each *lemma*. A *lemma* is a linguistic term for a group of inflected forms: arm, arms; serve, served, serves, serving.

How many word senses?

Words (lemmas) may have *many* more than two senses.
For example, serve:

- [FUNCTION]: The tree stump served as a table

How many word senses?

Words (lemmas) may have *many* more than two senses.
For example, serve:

- [FUNCTION]: The tree stump served as a table
- [ENABLE]: His evasive replies only served to heighten suspicion

How many word senses?

Words (lemmas) may have *many* more than two senses.
For example, serve:

- [FUNCTION]: The tree stump served as a table
- [ENABLE]: His evasive replies only served to heighten suspicion
- [DISH]: We serve only the rawest fish here

How many word senses?

Words (lemmas) may have *many* more than two senses.
For example, serve:

- [FUNCTION]: The tree stump served as a table
- [ENABLE]: His evasive replies only served to heighten suspicion
- [DISH]: We serve only the rawest fish here
- [ENLIST]: She served her country in the marines

How many word senses?

Words (lemmas) may have *many* more than two senses.
For example, serve:

- [FUNCTION]: The tree stump served as a table
- [ENABLE]: His evasive replies only served to heighten suspicion
- [DISH]: We serve only the rawest fish here
- [ENLIST]: She served her country in the marines
- [JAIL]: He served six years in Alcatraz

How many word senses?

Words (lemmas) may have *many* more than two senses.
For example, serve:

- [FUNCTION]: The tree stump served as a table
- [ENABLE]: His evasive replies only served to heighten suspicion
- [DISH]: We serve only the rawest fish here
- [ENLIST]: She served her country in the marines
- [JAIL]: He served six years in Alcatraz
- [TENNIS]: Nobody can return his double-reverse spin serve

How many word senses?

Words (lemmas) may have *many* more than two senses.
For example, serve:

- [FUNCTION]: The tree stump served as a table
- [ENABLE]: His evasive replies only served to heighten suspicion
- [DISH]: We serve only the rawest fish here
- [ENLIST]: She served her country in the marines
- [JAIL]: He served six years in Alcatraz
- [TENNIS]: Nobody can return his double-reverse spin serve
- [LEGAL]: They were served with subpoenas

How many word senses?

Words (lemmas) may have *many* more than two senses.
For example, serve:

- [FUNCTION]: The tree stump served as a table
- [ENABLE]: His evasive replies only served to heighten suspicion
- [DISH]: We serve only the rawest fish here
- [ENLIST]: She served her country in the marines
- [JAIL]: He served six years in Alcatraz
- [TENNIS]: Nobody can return his double-reverse spin serve
- [LEGAL]: They were served with subpoenas
- more?

How many word senses?

How can we test that these senses are really different?
We can construct a *zeugma*, which combines antagonistic senses in an uncomfortable way:

- Which flight serves breakfast?

How many word senses?

How can we test that these senses are really different?
We can construct a *zeugma*, which combines antagonistic senses in an uncomfortable way:

- Which flight serves breakfast?
- Which flights serve Tuscon?

How many word senses?

How can we test that these senses are really different?
We can construct a *zeugma*, which combines antagonistic senses in an uncomfortable way:

- Which flight serves breakfast?
- Which flights serve Tuscon?
- *Which flights serve breakfast and Tuscon?

How many word senses?

How can we test that these senses are really different?
We can construct a *zeugma*, which combines antagonistic senses in an uncomfortable way:

- Which flight serves breakfast?
- Which flights serve Tuscon?
- *Which flights serve breakfast and Tuscon?

How many word senses?

How can we test that these senses are really different?
We can construct a *zeugma*, which combines antagonistic senses in an uncomfortable way:

- Which flight serves breakfast?
- Which flights serve Tuscon?
- *Which flights serve breakfast and Tuscon?

(the asterisk is a linguistic notation for utterances which would not be judged to be grammatical by fluent speakers of a language)

How many word senses?

Sometimes the distinctions are not so clear. It's easy to differentiate:

- the muddy banks of the mighty mississippi
- i rob banks because that's where the money is

How many word senses?

Sometimes the distinctions are not so clear. It's easy to differentiate:

- the muddy banks of the mighty mississippi
- i rob banks because that's where the money is

But what about blood banks?
Is this a different sense from a financial bank? Let's try a zeugma:

- Vampire City has two kinds of banks: blood and money

The WSD task: Output

- What should the output of WSD be? What are the possible senses for each word?
- We could just look in the dictionary.

[1]**plunge** *verb* \ˈplənj\

plunged | **plung·ing**

Definition of PLUNGE

transitive verb

1 : to cause to penetrate or enter quickly and forcibly into something <*plunged* the dagger>

2 : to cause to enter a state or course of action usually suddenly, unexpectedly, or violently <*plunged* the nation into economic depression>

intransitive verb

1 : to thrust or cast oneself into or as if into water

2 a : to become pitched or thrown headlong or violently forward and downward; *also* **:** to move oneself in such a manner <*plunged* off the embankment>

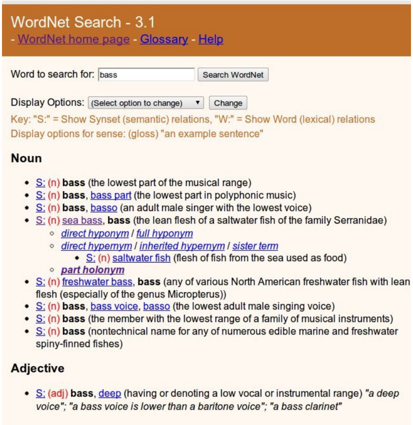
b : to act with reckless haste : enter suddenly or unexpectedly <*plunges* into project after project>

c : to bet or gamble heavily and recklessly

3 : to descend or dip suddenly <the stock's value *plunged*>

WordNet

WSD research is dominated by a computational resource called WORDNET. (`http://wordnet.princeton.edu`)



WordNet

- WordNet consists of roughly 100K *synsets*,
  groups of words or phrases with an identical meaning.
  (e.g., $\{\textsc{chump}^1, \textsc{fool}^2, \textsc{sucker}^1, \textsc{mark}^9\}$)
  A lemma is polysemous if it participates in multiple synsets.

WordNet

- WordNet consists of roughly 100K *synsets*, groups of words or phrases with an identical meaning. (e.g., $\{\text{CHUMP}^1, \text{FOOL}^2, \text{SUCKER}^1, \text{MARK}^9\}$) A lemma is polysemous if it participates in multiple synsets.
- WordNet also describes many other lexical relationships:
  - antonymy (x means the opposite of y)
  - hyponymy (x is a hyponym of y if x is-a y)
  - ...

WordNet

Some statistics of English Wordnet 3:

POS	polysemy
NOUN	1.24
VERB	2.17
ADJECTIVE	1.40
ADVERB	1.25

WordNet Senses: Pros and cons

- WordNet played a big role in helping WSD move from toy systems to to large-scale quantitative evaluations.

WordNet Senses: Pros and cons

- WordNet played a big role in helping WSD move from toy systems to to large-scale quantitative evaluations.
- WordNet's sense granularity may be too fine [IW06].
  Humans agree on 75-80% of examples using WordNet senses.

WordNet Senses: Pros and cons

- WordNet played a big role in helping WSD move from toy systems to to large-scale quantitative evaluations.
- WordNet's sense granularity may be too fine [IW06].
  Humans agree on 75-80% of examples using WordNet senses.
- Are word senses real?
  The premise that word senses can be differentiated in a task-neutral way has been criticized as linguistically naïve [Kil97].

WordNet Senses: Pros and cons

- WordNet played a big role in helping WSD move from toy systems to to large-scale quantitative evaluations.
- WordNet's sense granularity may be too fine [IW06].
  Humans agree on 75-80% of examples using WordNet senses.
- Are word senses real?
  The premise that word senses can be differentiated in a task-neutral way has been criticized as linguistically naïve [Kil97].
- WordNets are heavyweight.
  - expensive to develop for new languages
  - become outdated as language changes
    (consider: I'm dead tired, sick as a positive adjective, etc)
  - Would WordNet have good coverage for Twitter?

Translation Sets as Word Senses

An alternative is to use translation to differentiate word senses.

- Since bill is translated as pico or cuenta in spanish, there are clearly two senses.

Translation Sets as Word Senses

An alternative is to use translation to differentiate word senses.

- Since bill is translated as pico or cuenta in spanish, there are clearly two senses.
- But if there is no language with different spellings of the purported senses, then they are not meaningfully different. Possible example:

Translation Sets as Word Senses

An alternative is to use translation to differentiate word senses.

- Since bill is translated as pico or cuenta in spanish, there are clearly two senses.
- But if there is no language with different spellings of the purported senses, then they are not meaningfully different. Possible example:
  - engine^4 (MACHINE FOR CONVERTING ENERGY INTO FORCE)

Translation Sets as Word Senses

An alternative is to use translation to differentiate word senses.

- Since bill is translated as pico or cuenta in spanish, there are clearly two senses.
- But if there is no language with different spellings of the purported senses, then they are not meaningfully different. Possible example:
  - engine^4 (MACHINE FOR CONVERTING ENERGY INTO FORCE)
  - engine^6 (COMPUTER PROGRAM THAT PERFORMS A FUNDAMENTAL FUNCTION)

Translation Sets as Word Senses

An alternative is to use translation to differentiate word senses.

- Since bill is translated as pico or cuenta in spanish, there are clearly two senses.
- But if there is no language with different spellings of the purported senses, then they are not meaningfully different. Possible example:
    - engine^4 (MACHINE FOR CONVERTING ENERGY INTO FORCE)
    - engine^6 (COMPUTER PROGRAM THAT PERFORMS A FUNDAMENTAL FUNCTION)

Translation Sets as Word Senses

An alternative is to use translation to differentiate word senses.

- Since bill is translated as pico or cuenta in spanish, there are clearly two senses.
- But if there is no language with different spellings of the purported senses, then they are not meaningfully different. Possible example:
  - engine^4 (MACHINE FOR CONVERTING ENERGY INTO FORCE)
  - engine^6 (COMPUTER PROGRAM THAT PERFORMS A FUNDAMENTAL FUNCTION)

Most WSD research has focused on WordNet, so we will too.

WSD Tasks

- **Synthetic** data: different words are conflated (banana-phone), the system must identify the original word.

WSD Tasks

- **Synthetic** data: different words are conflated (banana-phone), the system must identify the original word.
- **Lexical sample**: disambiguate a few target words (e.g., "plant" etc).

  This was the task of the first large-scale WSD evaluation, SENSEVAL-1 (1998).

WSD Tasks

- **Synthetic** data: different words are conflated (banana-phone), the system must identify the original word.
- **Lexical sample**: disambiguate a few target words (e.g., "plant" etc).

  This was the task of the first large-scale WSD evaluation, SENSEVAL-1 (1998).
- **All-words** WSD: a sense must be identified for every token.

WSD Tasks

- **Synthetic** data: different words are conflated (banana-phone), the system must identify the original word.
- **Lexical sample**: disambiguate a few target words (e.g., "plant" etc).

  This was the task of the first large-scale WSD evaluation, SENSEVAL-1 (1998).
- **All-words** WSD: a sense must be identified for every token.
  - A **semantic concordance** is a corpus in which each open-class word (nouns, verbs, adjectives, and adverbs) is tagged with its word sense from the target dictionary or thesaurus.

WSD Tasks

- **Synthetic** data: different words are conflated (banana-phone), the system must identify the original word.
- **Lexical sample**: disambiguate a few target words (e.g., "plant" etc).
  This was the task of the first large-scale WSD evaluation, SENSEVAL-1 (1998).
- **All-words** WSD: a sense must be identified for every token.
  - A **semantic concordance** is a corpus in which each open-class word (nouns, verbs, adjectives, and adverbs) is tagged with its word sense from the target dictionary or thesaurus.
  - SemCor is a semantic concordance built from 234K tokens of the Brown corpus.
    As of Sunday$_n^1$ night$_n^1$ there was$_v^4$ no word$_n^2$ ...

How can we solve WSD?

- How can we tell living plants from manufacturing plants?

How can we solve WSD?

- How can we tell living plants from manufacturing plants?
- **Context**

How can we solve WSD?

- How can we tell living plants from manufacturing plants?
- **Context**
  - Town officials are hoping to attract new manufacturing plants through weakened environmental regulations.
  - The endangered plant plays an important role in the local ecosystem.

How can we solve WSD?

- How can we tell living plants from manufacturing plants?
- **Context**
  - Town officials are hoping to attract new manufacturing plants through weakened environmental regulations.
  - The endangered plant plays an important role in the local ecosystem.
- Approaches:
  - Knowledge-based
  - Supervised
  - Semi-supervised
  - Unsupervised

The Lesk Algorithm

- For each sentence s_i and target word w_{ij}
  - Set $maxOverlap \leftarrow 0$, $bestSense \leftarrow \varnothing$
  - For each possible sense t
    - Compute word overlap between s_i and definition $w_{ij}[t]$
    - If greater than $maxOverlap$, then update $maxOverlap$ and $bestSense$.

The Lesk Algorithm

- For each sentence s_i and target word w_{ij}
  - Set $maxOverlap \leftarrow 0$, $bestSense \leftarrow \emptyset$
  - For each possible sense t
    - Compute word overlap between s_i and definition $w_{ij}[t]$
    - If greater than *maxOverlap*, then update *maxOverlap* and *bestSense*.

Example text: I stopped by the **bank** to try to get a loan
Example definitions:

- Bank^1: financial institution which borrows and loans money
- Bank^2: body of land adjacent to a river

The first sense is preferred because the word "loan" appears in both the definition and the query sentence.

Corpus Lesk

"Corpus Lesk" weights each word in the context with its (log) inverse document frequency:

$$idf_i = \log\left(\frac{\#|d|}{\#|\{d : w_i \in d\}|}\right),$$

where $\#|d|$ is the number of documents in the dataset, and $\#|\{d : w_i \in d\}|$ is the number which contain word w_i.

- Little reward for a match of a common word like the, more reward from a relatively rare word like mortgage.
- IDF-weighting is a good trick in many information retrieval and NLP tasks.

Selectional restrictions

Some verbs have strong selectional restrictions about their arguments:

- They closed the bank^1 after discovering its malfeasance.
- They rested on the bank^2 of the Seine.
- Closed can only take an argument which is an ORGANIZATION.
- Rested can only take an argument which is a PHYSICAL-OBJECT.

Some ontologies categorize common nouns in terms of such properties.

Selectional association

Explicit constraints are brittle, so Resnick (1997) introduced a softer probabilistic approach.
Preliminaries:

- $P(c)$ is the overall corpus probability of class c (e.g., LEGAL DOCUMENTS)
- $P(c|v)$ is the probability of class c as an argument of verb v

Selectional association

Explicit constraints are brittle, so Resnick (1997) introduced a softer probabilistic approach.
Preliminaries:

- $P(c)$ is the overall corpus probability of class c (e.g., LEGAL DOCUMENTS)
- $P(c|v)$ is the probability of class c as an argument of verb v
- $D_{KL}(P_1||P_2)$ is the Kullback-Liebler (KL) divergence between the probability distributions P_1 and P_2.
  $D_{KL}(P_1||P_2) = \sum_x P_1(x) \log \frac{P_1(x)}{P_2(x)}$.
  - $D_{KL}(P_1||P_2) \geq 0$
  - $D_{KL}(P_1||P_2) = 0$ implies $P_1 = P_2$.
  - In general, $D_{KL}(P_1||P_2) \neq D_{KL}(P_1||P_2)$

Selectional association

The *information content* of verb v is the KL-divergence of $P(C|v)$ with the prior distribution $P(C)$.

$$
\begin{aligned}
S_R(v) =& D_{KL}(P(C|v)||P(C)) \\
=& \sum_{c \in C} P(c|v) \log \frac{P(c|v)}{P(c)}
\end{aligned}
$$

Selectional association

The *information content* of verb v is the KL-divergence of $P(C|v)$ with the prior distribution $P(C)$.

$$\begin{aligned} S_R(v) =& D_{KL}(P(C|v)||P(C)) \\ =& \sum_{c \in C} P(c|v) \log \frac{P(c|v)}{P(c)} \end{aligned}$$

The selectional association of a verb and a particular class can be measured by the ratio:

$$A_R(v,c) = \frac{P(c|v) \log \frac{P(c|v)}{P(c)}}{S_R(v)} = \frac{P(c|v) \log \frac{P(c|v)}{P(c)}}{\sum_{c'} P(c'|v) \log \frac{P(c'|v)}{P(c')}}$$

We choose the sense whose class has the highest selectional association with the verb on which it depends.

Supervised WSD

- With labeled data, we can treat WSD as a standard supervised learning problem.
- Some features
  - Bag-of-words
  - Positional (collocation) features
  - Patterns
  - Syntax
  - Document features

Bag-of-words features

Bag-of-words models are a very typical approach. For example,

$$f(y, \text{bank}, \text{I went to the bank to deposit my paycheck}) = \{\langle \text{went}, y\rangle : 1, \langle \text{deposit}, y\rangle : 1, \langle \text{paycheck}, y\rangle : 1\}$$

Bag-of-words features

Bag-of-words models are a very typical approach. For example,

$$f(y, \text{bank}, \text{I went to the bank to deposit my paycheck}) = \{\langle \text{went}, y\rangle : 1, \langle \text{deposit}, y\rangle : 1, \langle \text{paycheck}, y\rangle : 1\}$$

Some examples (Mihalcea and Pederson 2006)

- bank[FINANCIAL]:

  a an and are ATM Bonnie card charges check Clyde criminals deposit famous for get I much My new overdraft really robbers the they think to too two went were

- bank[RIVER]:

  a an and big campus cant catfish East got grandfather great has his I in is Minnesota Mississippi muddy My of on planted pole pretty right River The the there University walk Wets

Positional (collocation) features

- An extension of bag-of-words models is to encode the position of each context word, e.g.,

  $$f(y, \text{bank}, \text{I went to the bank to deposit my paycheck}) = \{\langle i-3, \text{went}, y\rangle : 1, \langle i+2, \text{deposit}, y\rangle : 1, \langle i+4, \text{paycheck}, y\rangle : 1\}$$

- J&M (optional textbook) call these collocation features; the POS tag of each word can also be included.

Pattern features

Pattern features extend the idea of positional features with explicit, regex-like patterns:

- bank account
- bank of COUNTRY.

Such features are often used in combination with non-linear classifiers such as decision lists.

Syntactic features

- Rather than look at local neighbors, we can give special priority to the heads of phrases.
- For example, in

  $$\textit{I deposited my paycheck when I got to the bank},$$

  the most revealing features are deposit and paycheck.
- deposit is the head of the main verb phrase for the sentence, and paycheck is the direct object.
- This is a clue that they are more relevant than the words immediately surrounding bank.

Document-level features

According to the "one-sense-per-discourse" heuristic, a document about financial institutions is very unlikely to use the word bank in the river bank sense.

Word	Senses	Accuracy	Applicblty
plant	living/factory	99.8 %	72.8 %
tank	vehicle/contnr	99.6 %	50.5 %
poach	steal/boil	100.0 %	44.4 %
palm	tree/hand	99.8 %	38.5 %
axes	grid/tools	100.0 %	35.5 %
sake	benefit/drink	100.0 %	33.7 %
bass	fish/music	100.0 %	58.8 %
space	volume/outer	99.2 %	67.7 %
motion	legal/physical	99.9 %	49.8 %
crane	bird/machine	100.0 %	49.1 %
Average		99.8 %	50.1 %

(Yarowsky 1995)

Decision Lists

Early work focused on *decision lists*. They're not used much now, but the method is illustrative (Yarowsky, 1994).

- Features:
  - words immediate to the left or right
  - pairs of words to the left or right
  - words within k positions of the target ($10 \leq k \leq 50$)
- Algorithm: score each example by the max log-odds ratio

$$score(i) = \max_f \log \frac{Pr(sense = i|f(w))}{\sum_{j \neq i} Pr(sense = j|f(w))}$$

- Rather than combining many different features, just take the one with the strongest opinion!
- If you tune the smoothing carefully, this can work.
  In fact, it got the best score at the first SensEval (1998).

Classification approaches

- Naive Bayes (can you see how to do this?)
- Discriminative classifiers (tuesday): logistic regression, SVM
- "Ensemble methods" which combine different types of classifiers

Is Word Sense Disambiguation Important?

- Early machine translation researchers were really worried about WSD.
  - bill[BIRD JAW] → pico
  - bill[INVOICE] → cuenta

Is Word Sense Disambiguation Important?

- Early machine translation researchers were really worried about WSD.
  - bill[BIRD JAW] $\rightarrow$ pico
  - bill[INVOICE] $\rightarrow$ cuenta
- Bar-Hillel, an expert-turned-skeptic, poses this problem:

  "Little John was looking for his toy box. Finally he found it. The box was in the pen." Is pen a writing instrument or a place where children play?

Is Word Sense Disambiguation Important?

- Early machine translation researchers were really worried about WSD.
  - bill[BIRD JAW] $\rightarrow$ pico
  - bill[INVOICE] $\rightarrow$ cuenta
- Bar-Hillel, an expert-turned-skeptic, poses this problem:

  > *"Little John was looking for his toy box. Finally he found it. The box was in the pen." Is pen a writing instrument or a place where children play?*

- The suggestion is this example requires deep knowledge and inference (a box is bigger than a pen[WRITING], but not bigger than a pen[ENCLOSURE]).
- Bar-Hillel was so discouraged that he gave up on MT!

The Role of WSD Today

- WSD was also thought to be important for information retrieval: bass experts, help with cures, etc.
- Many thought the NLP pipeline required a WSD module.
  preprocessing → POS tagging → WSD → application

[1] The survey goes on to argue that WSD will become more relevant as performance improves.

The Role of WSD Today

- WSD was also thought to be important for information retrieval: bass experts, help with cures, etc.
- Many thought the NLP pipeline required a WSD module. preprocessing → POS tagging → WSD → application
- However, years of research on WSD have produced little evidence that it helps downstream applications. A recent survey of WSD notes:

  > *Unfortunately, to date explicit WSD has not yet demonstrated real benefits in human language technology applications (Navigli 2009).*[1]

[1] The survey goes on to argue that WSD will become more relevant as performance improves.

The Role of WSD Today

- There is some recent evidence that WSD helps translation (Chan, Ng and Chiang 2007; Carpuat and Wu 2007)
- But in many tasks, higher-order n-grams encode much the same information as WSD.
  - If we have the bigram bank teller as a feature, we don't need to disambiguate bank.
  - Phrase-based machine translation uses a similar idea.

Homework 2

- Download the SemCor data.
- Compare the word sense annotations with WordNet online.
- Explain why alternative senses were not chosen.
- Do word sense annotations for one sentence of text from an (English language) blog that you like.

N. Ide and Y. Wilks.
Making sense about sense, 2006.

Adam Kilgarriff.
"i don't believe in word senses".
CoRR, cmp-lg/9712006, 1997.