

Words, Concepts, Meanings

July 11, 2011

LSA Institute, Boulder CO

© Christiane Fellbaum

Outline

- What are we trying to “teach” a computer? What are the challenges and how do we approach them? How well are we doing?
- Lexical Polysemy/Homonymy
- Zipf's Law
- What's a sense? Words and Concepts
- Word and World Knowledge
- Need for Word Sense Disambiguation in NLP
- Approaches: Supervised, unsupervised learning
- Clustering
- Reference Lexical Resources
- Lexical-semantic patterns for induction, verification of relations
- Holy Grail for supervised learning: Semantic Concordance
- Are senses discrete?

Some Definitions

- The **Lexicon**: the component of grammar that includes **knowledge of words**
- Very large (> 40 000 words) Do the dictionary test!
- Open-ended
- Dynamic--never static
- Acquisition is on-going, lifelong (so is loss)

Knowledge of words

- Sound (pronunciation)
- Meaning/concept behind the word
- Syntax (e.g., argument selection for verbs)
- Semantic roles (is the subject of the verb *feel* an Agent or an Experiencer or...?)
- Morphology (e.g., plural formation)
- Selectional restrictions/preferences:
 - strong vs. ?powerful tea*
 - dog house vs. ?canine house/domicile*
- Some languages: class membership marker
- written representation

Key question:

How much of this knowledge must we get computers to “understand” and how do we do it?

Word vs. world knowledge

Lexical vs. “encyclopedic” knowledge

Fuzzy boundary. Example: “radon”

World dictionary:

a colourless radioactive element of the rare gas group, the most stable isotope of which, radon-222, is a decay product of radium. It is used as an alpha particle source in radiotherapy. Symbol: Rn; atomic no: 86; half-life of 222 Rn: 3.82 days; valency: 0; density: 9.73 kg/m³; melting pt: --71°C; boiling pt: --61.7°C

Wikipedia:

Radon (play /'reɪdɒn/ ray-don) is a chemical element with symbol Rn and atomic number 86. It is a radioactive, colorless, odorless, tasteless noble gas, occurring naturally as the decay product of uranium. It is one of the densest substances that remains a gas under normal conditions and is considered to be a health hazard due to its radioactivity. Its most stable isotope, 222Rn, has a half-life of 3.8 days. Due to its intense radioactivity, it has been less well-studied by chemists, but a few compounds are known...

Some definitions: **Word**

Intuitively easy, but...

(1) Word = concept/unit of meaning?

Nootka verb (Sapir):

I have been accustomed to eating twenty round objects while engaging in...

Word

(2) Word = sequence of letters between white empty spaces

Not all languages use letters (~phonemic representation) or segment words

Word

(3) Word = minimal free morpheme

Status of bound morphemes (*un-*, *-ed*)?

Compounds (*Miami-to-Montreal* train)?

Multi-Word Units (*put up*, *end-of-life care*)

idioms (*hit the ceiling*, *won't hear of it*, *blow one's stack*)

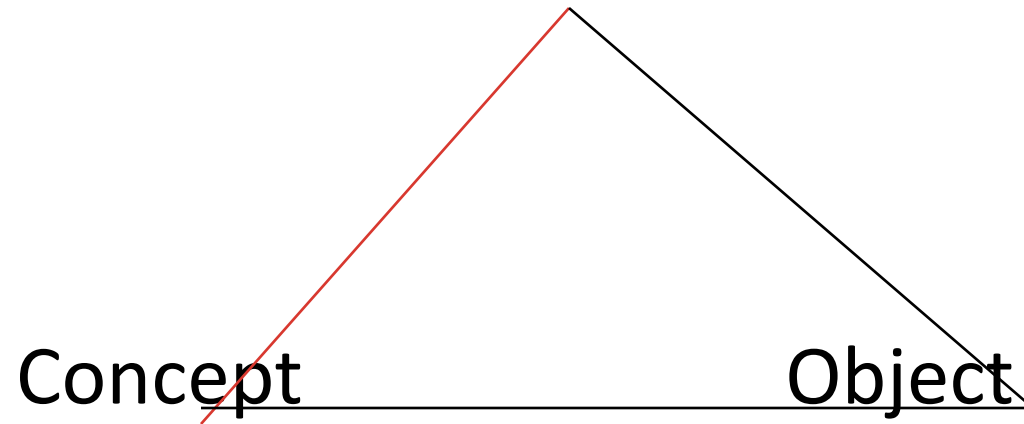
Non-compositional MWUs: *hot dog*, *hothouse*

Words = form-meaning mappings

A word form is often called a lemma (pl. lemmas or lemmata)

Semiotic triangle

Word



Words

Two important properties of words
(universal?)

Polysemy/Homonymy

Synonymy

The lexicon

Repository of all form-meaning pairs

Ideal (?): one form, one meaning:

Every word form has exactly one meaning

Every meaning is expressible by exactly one
word form

Real Life: (mostly) many-to-many mappings

Synonymy

One meaning/concept is expressed by several different word forms:

{beat, hit, strike}, {shut, close}

{car, motorcar, auto, automobile}

{big, large}, {difficult, hard}

Polysemy

One word form expresses multiple meanings

{*table*, tabular_array}

{*table*, piece_of_furniture}

{*table*, mesa}

{*table*, postpone}

Polysemy, homonymy

Homonymy: unrelated meanings:

pitch (acoustic property/tar)

bar (saloon/stick)

bass(fish, musical instrument)

Polysemy: related meanings:

bass (vocal range, singer)

Borderline can be fuzzy.

Polysemy test

Zeugma:

?He left town and the bills on the table

leave₁: go away from a place

leave₂: leave behind, cause to remain in a specific place

cf.:

He left the town and, later, the country

He left the bills and a bottle of beer on the table

Polysemy

Regular, systematic polysemy:

book, newspaper,... (object, content)

chicken, beans,...(animal/plant, food/dish)

Metonymy (part-whole polysemy)

The House of Saud issued a statement

The office isn't answering the phone

The lexical matrix

A “map” of the lexicon

Plotting polysemy and synonymy

Most cells are empty

| | | | | | |
|-------|-----|-----------|-------|---------|------|
| pitch | tar | | | | |
| pitch | | frequency | | | |
| pitch | | | throw | | toss |
| | tar | | | tarball | |
| | | | | | |
| | | | | | |

Synonymy, polysemy

perfect for poets: great power of expression
(synonymy) and deliberate ambiguity or
vagueness (polysemy)

Terrible for computers!

Zipf's Law

A power law:

Frequency of a word is inversely proportional to its rank in a frequency table

Most frequent word occurs twice as often as the second most frequent word

Three times as often as the third most frequent word, etc.

This is not quite true...

Frequency rank lists and frequencies

| Brown Corpus (1 mio words) | British National Corpus (100 mio) |
|----------------------------|-----------------------------------|
| the 69975 | the 6055159 |
| be 39175 | be 654445 (all inflected forms) |
| of 36432 | of 3051609 |
| and 28872 | and 2632194 |
| a 23073 | a 2168817 |
| in 20870 | in 1944328 |

(Zipf's Law holds for the middle of the curve; there's a very long tail with many words that are used only once in a corpus)

Open questions about Zipf's law

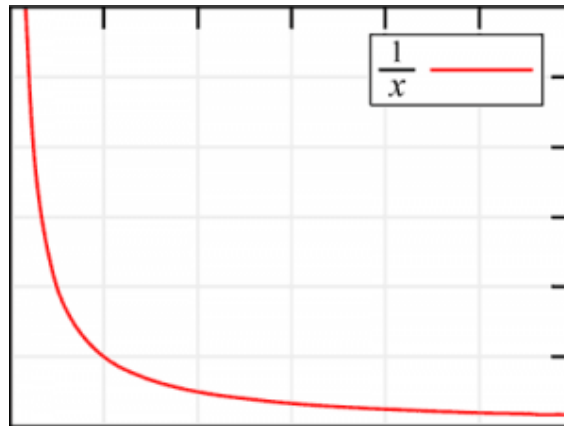
Zipf claimed it holds for James Joyce's *Ulysses*

Can't be reproduced...could it be true for other texts? Why (not)?

How would ZL apply to other languages?

To make things worse for WSD...

The most frequent word forms are the most polysemous



Polysemy Problem

The Big Challenge for automatic language processing:

Which sense of a word in a context is the one intended by the speaker/writer?

Humans have no problem with disambiguation!

Example

Type *turkey* into your search engine

The string *turkey* can refer to a bird, the meat of the bird, a country, a foolish person

The search engine cannot guess which concept you have in mind! It returns pages for all meanings of *turkey*.

Example cont'd.

Give the search engine a better chance...

More specific query might eliminate ambiguity and irrelevant information:

“Do turkeys fly?”

Ask Jeeves returns pages for

(a) animal facts: Turkey

(b) cheap flights to Turkey

Still too much irrelevant information! Humans would have ruled out (b)

Synonymy

In response to the query

“Where can I find a lawyer in Kalamazoo?”

search engines return web pages with **lawyers** in Kalamazoo. But many relevant web pages are missed that refer to **attorneys**

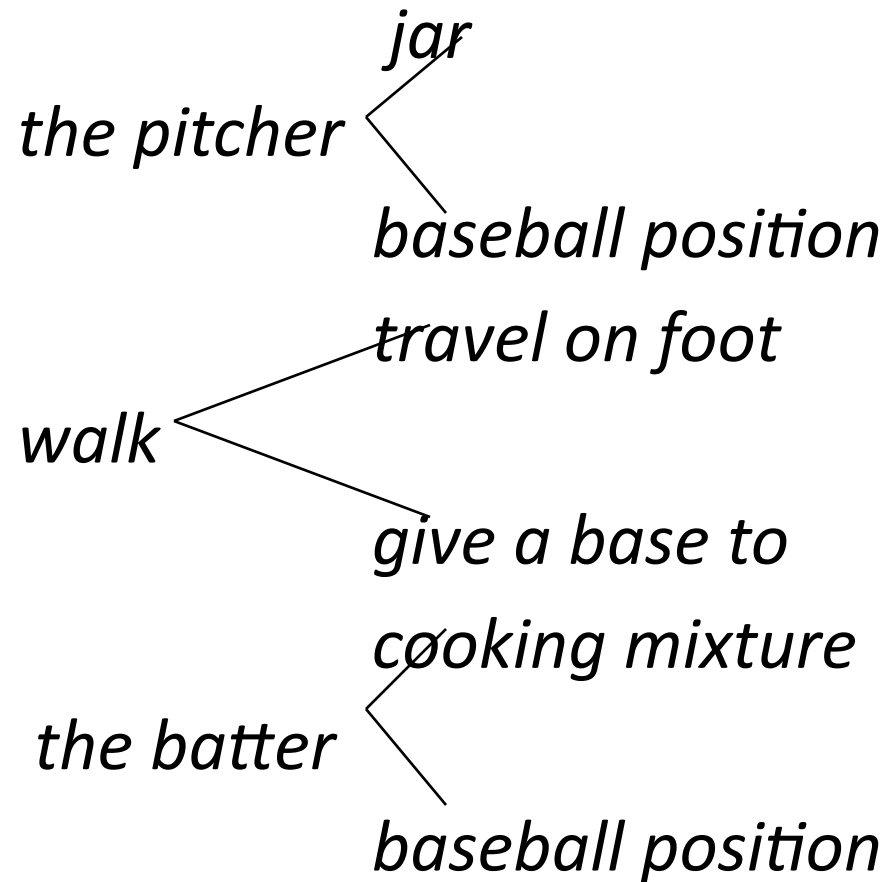
Really smart search engine would return pages covering lawyers and attorneys

Polysemy and synonymy

Polysemy can cause overgeneration of hits
(several concepts, including irrelevant ones,
associated with one string)

Synonymy can cause undergeneration (no hits
because not all strings that can express a given
concept were used in the query)

Automatic text understanding



Word sense disambiguation

WSD=map word form in a context to the appropriate word meaning

Major challenge!

Basic assumptions

Context is semantically coherent (*pace*
“colorless green ideas sleep furiously”)

Words in a context are semantically related
and disambiguate one another in a given
context

Meaning of a word is the sum of its possible
contexts

Children’s acquisition of word meaning
proceeds via contexts (not definitions!)

People have disambiguated words for a long time
E.g., modern dictionary making:
Lexicographers extract from a corpus KWIC lines
(Key Words In Context)

“You shall know a word by the company it keeps”

J. R. Firth (1957)

WSD seems straightforward

From the British National Corpus, 100 mio words:

*it is then useless for the moth to flee, because it will
have been detected and the bat can fly faster than it.*

depending upon how starved a bat is.

was ideal for getting the feel of bat on ball

*how he killed a stranger with a single blow from a
baseball bat.*

but isn't

*But the old **bat** was always hanging around — and then the next thing I knew she was telling me you'd signed the lease.*

*In 1895 he began to build his first full-size glider, the **Bat***

*A **bat** is a machine, whose internal electronics are so wired up that its wing muscles cause it to home in on insects, as an unconscious guided missile homes in on an aeroplane.*

*For the black **bat**, night, has flown*

- Mapping tokens (specific occurrences of a word form) to a dictionary entry is difficult
- “Discovering” (inducing) words senses and sense distinctions is hard, too

Machine Learning

Automatic systems can “learn” to

- discriminate senses (distinguish different senses)
- disambiguate words (against a reference/standard, such as a dictionary)

Machine Learning

(Semi-)supervised learning

System trains on hand-annotated data (“gold standard”)

Unsupervised learning

makes no use of human-annotated data

Supervised WSD

Mimics human disambiguation

Requires a training corpus with manually annotated (labeled) data

Humans read text, link each (content) word to the context-appropriate entry in a lexicon

Most commonly used: SemCor, GlossCor (both annotated against Princeton WordNet*)

SemCor: part of the Brown Corpus

GlossCor: many of the glosses (definitions) accompanying WordNet senses

*a digital dictionary, more later

Supervised WSD

System “trains” on training corpus with labeled data

Builds a classifier based that assign labels

Classifier is built by extracting features associated with
(and predictive of) specific senses

Finally, system performs WSD on (unlabeled) “testing”
corpus

Supervised WSD

- Corpus is pre-processed (parsed, lemmatized, POS tagged)
- Context features are extracted from a window of n words to the right and left of the target word (n-gram)

Information about the context words is encoded in a vector

Vectors are the input to the learning system

Context features

Collocational: encode specific position relative to target word (in addition to other information)

Typical features are the context words, their POS, stemmed forms

Good for local contexts

Example

Collocational feature vector extracted from a window of 2 words on each side of the target

Consists of with words and their POS:

*and the **bat** can fly faster than it*

[and,CC, the,Det,can,Mod,fly,VB]

Context features

bag-of-words:

Features consist of unordered neighboring words within a given window

For a given target word, feature set may consist of ten most frequently found context words

{*bat*₁: *fly, faster, starved,...*}

{*bat*₂: *ball, blow, baseball,...*}

Represent word meaning as a set of contexts

For a given token of that word, a vector shows whether or not these words occur in the context (binary: 0,1)

Good for larger (“global”, discourse) contexts

Most systems use both contextual and bag-of-words vectors

Evaluating the system

Check agreement with manually assigned labels,
which are not shown to the system during
testing

- Vectors for different tokens (specific occurrences of the target word) will differ
- Experimenter must define which vectors identify correct targets
- Modify feature set?

Minimally/semi-supervised learning

- Bootstrapping:
- System “discovers” new senses based on previously learned senses
- Doesn’t just match new tokens on previously learned senses
- Needs small set of labeled data (“seeds”)
- Seeds are either manually labeled or assumed (e.g., *bass* in the context of *play* is assumed to refer to the musical rather than the fish sense; Yarowsky 1993)

- System learns a classifier based on labeled data
- Applies it to unlabeled data
- Data with highest “confidence” rating is added to the training data
- Classifier is refined, new data are labeled
- Etc.

Unsupervised learning

- No manually labeled training data
- No use of sense inventory defined by humans
- System “discovers” different senses automatically
- Problem: system discriminates senses but doesn’t “know” the meanings/contents of the different senses
- Basically just clusters

Unsupervised WS learning

Represent each token of a target word in the training corpus by a set of distributional context vectors

Feature vectors: neighboring words (n-gram), neighbors in specific grammatical relations (object, subject of verb, etc.--needs parsed corpus)

Feature f for word w :

$$f = (r, w')$$

w' = related word

r = relation

e.g., (*object-of, eat*) could be a feature of (one sense of) *apple*

Frequency/probability gives weights to features (i.e., a feature that is found more often “carries greater weight” in determining the correct sense)

Probability of finding a feature f given the target word w (probability as a measure of association):

$$P(f | w) = \frac{\text{count}(f, w)}{\text{count}(w)}$$

Mutual association

- But probability/frequency alone isn't all that useful
- Stop words (*the, a, my, on*) are frequent but not informative
- Which (content) words collocate?
- Measure of association: mutual information (Church & Hanks 1989)
- Pointwise mutual information (based on Fano 1961)

Association measure

How often do a word and feature co-occur ($P(w,f)$) compared to their expected independent occurrence ($P(w)P(f)$)?

Pointwise Mutual Information between target word and a feature

$$PMI(w,f) = \log_2 \frac{P(w,f)}{P(w)P(f)}$$

Can build “thesaurus” automatically by clustering semantically similar words without specifying meanings of word forms or their relation to other words, as in Roget’s

Success?

Supervised methods work best

For homographs (*bat, pitch, check, bar*), learning systems perform at about 90% correct (relative to human judgments)

For polysemes, agreement with human judgment is only 55-70%

Depends partly on grainedness of reference sense inventory (i.e., identifying the most appropriate sense of a word with 20 senses is harder than for words with 5 senses; both for humans and machines more later)

Knowledge-Based Approaches

Determine which sense in a lexical resource (like WordNet) a given token (occurrence) of a word represents

“Dumb” method: assume that the most frequent* sense of a polysemous word is the context-appropriate one

Works amazingly well: 65-70% correct for nouns

*Frequency=how often a given sense has been selected in a manual corpus annotation task...and herein lies another problem...to be discussed

We want to do better...