Introduction to NLP
Lexical Semantics[1]

Alexandre Rademaker

FGV/EMAp

October 20, 2021

Outline of today's lecture

Lexical semantics: semantic relations

Polysemy

Word sense disambiguation

Grounding

Lexical semantics

- ► Limited domain: mapping to some knowledge base term(s). Knowledge base constrains possible meanings.
- ► Issues for broad coverage systems:
  - ► Boundary between lexical meaning and world knowledge.
  - ► Representing lexical meaning.
  - ► Acquiring representations.
  - ► Polysemy and multiword expressions.

Approaches to lexical meaning

- ▶ Formal semantics: extension — what words denote (e.g., cat′: the set of all cats).

Approaches to lexical meaning

▶ Formal semantics: extension — what words denote (e.g., cat′: the set of all cats). But . . .

Approaches to lexical meaning

- Formal semantics: extension — what words denote (e.g., cat′: the set of all cats). But ...
- Semantic primitives: e.g., *kill* means CAUSE (NOT (ALIVE)).

Approaches to lexical meaning

- Formal semantics: extension — what words denote (e.g., cat′: the set of all cats). But . . .
- Semantic primitives: e.g., *kill* means CAUSE (NOT (ALIVE)). But . . .

Approaches to lexical meaning

- Formal semantics: extension — what words denote (e.g., cat′: the set of all cats). But . . .
- Semantic primitives: e.g., *kill* means CAUSE (NOT (ALIVE)). But . . .
- Meaning postulates:

  $$\forall e, x, y[\text{kill}'(e, x, y) \to \exists e'[\text{cause}'(e, x, e') \land \text{die}'(e', y)]]$$

Approaches to lexical meaning

- Formal semantics: extension — what words denote (e.g., cat': the set of all cats). But ...
- Semantic primitives: e.g., *kill* means CAUSE (NOT (ALIVE)). But ...
- Meaning postulates:

  $$\forall e, x, y[\text{kill}'(e, x, y) \rightarrow \exists e'[\text{cause}'(e, x, e') \wedge \text{die}'(e', y)]]$$

  But ...

Approaches to lexical meaning

- ▶ Formal semantics: extension — what words denote (e.g., cat′: the set of all cats). But . . .
- ▶ Semantic primitives: e.g., *kill* means CAUSE (NOT (ALIVE)). But . . .
- ▶ Meaning postulates:

  $$\forall e, x, y[\text{kill}'(e, x, y) \rightarrow \exists e'[\text{cause}'(e, x, e') \wedge \text{die}'(e', y)]]$$

  But . . .
- ▶ Ontological relationships: informal or formal (description logics): this lecture (informal approaches).
- ▶ Distributional approaches (word embeedings).

Examples to think about

- tomato
- table
- thought
- democracy
- push
- sticky

Hyponymy: IS-A

- (a sense of) *dog* is a hyponym of (a sense of) *animal*
- *animal* is a hypernym of *dog*
- hyponymy relationships form a taxonomy
- works best for concrete nouns

Some issues concerning hyponymy

- ▶ not useful for all words: *thought*, *democracy*, *push*, *sticky*?
- ▶ individuation differences: is *table* a hyponym of *furniture*?
- ▶ multiple inheritance: e.g., is *coin* a hyponym of both *metal* and *money*?
- ▶ what does the top of the hierarchy look like?

Other semantic relations

classical

Meronomy: PART-OF   e.g., *arm* is a meronym of *body*, *steering wheel* is a meronym of *car* (piece vs part)

Synonymy   e.g., *aubergine*/*eggplant*.

Antonymy   e.g., *big*/*little*

others

Near-synonymy/similarity   e.g., *exciting*/*thrilling*
            e.g., *slim*/*slender*/*thin*/*skinny*

More at https://globalwordnet.github.io/gwadoc/

WordNet

- large scale, open source resource for English
- hand-constructed
- wordnets being built for other languages
- organized into synsets: synonym sets (near-synonyms)
- Overview of adj red in `wn` command and Emacs Mode
- boy

Hyponymy in WordNet

```
Sense 6
big cat, cat
      => leopard, Panthera pardus
         => leopardess
         => panther
      => snow leopard, ounce, Panthera uncia
      => jaguar, panther, Panthera onca,
                                    Felis onca
      => lion, king of beasts, Panthera leo
         => lioness
         => lionet
      => tiger, Panthera tigris
         => Bengal tiger
         => tigress
```

Using hyponymy

- Semantic classification: e.g., for named entity recognition. e.g., JJ Thomson Avenue is a place.
- RTE style inference: find/discover
- Word sense disambiguation
- Query expansion in search

Collocation

- two or more words that occur together more often than expected by chance (informal description — there are others)
- some collocations are multiword expressions (MWE), green tea
- non-MWEs: heavy snow or *young boy*?
- what about geological fault?

Open Wordnet for Portuguese

- http://openwordnet-pt.org/
- coverage vs soundness
- can we find:
  - cargo – no sentido de posicao social,
  - pleito – no sentido de eleicao, nao de requisicao
  - posse – no sentido de dia da posse, cerimonia de inauguracao como deputado ou senador ou ministro
- glosstag corpus and its semantic representation
- corpus annotation

Polysemy

- ▶ homonymy: unrelated word senses. *bank* (raised land) vs *bank* (financial institution)
- ▶ *bank* (financial institution) vs *bank* (in a casino): related but distinct senses.
- ▶ *bank* (N) (raised land) vs *bank* (V) (to create some raised land): regular polysemy. Compare *pile*, *heap* etc
- ▶ vagueness: *bank* (river vs snow vs cloud)? see here

No clearcut distinctions.
Dictionaries are not consistent.

Word sense disambiguation

Needed for many applications, problematic for large domains.
Assumes that we have a standard set of word senses (e.g.,
WordNet)

- ► frequency: e.g., *diet*: the food sense (or senses) is much
  more frequent than the parliament sense (Diet of Wurms)
- ► collocations: e.g. *striped bass* (the fish) vs *bass guitar*:
  syntactically related or in a window of words (latter
  sometimes called 'cooccurrence'). Generally 'one sense
  per collocation'.
- ► selectional restrictions/preferences (e.g., *Kim eats bass*,
  must refer to fish

WSD techniques

- supervised learning: sense-tagged corpora are difficult to construct, algorithms need far more data than POS tagging. (google, datasets)
- unsupervised learning (pywsd, ukb)
- selectional preferences: don't work very well by themselves, useful in combination with other techniques

Standalone WSD

Once a very common research topic, now less studied:

- ▶ Evaluation issues
- ▶ Lack of a good standard

Grounding

- meaning isn't (just) about symbols: humans need to recognize and manipulate things in the world.
- 'grounding': relate symbols to the real world (often associated with Harnad, but other authors too).
- is grounding an essential part of meaning?
- preliminary/abstract discussion here

Turing: 'Computing machinery and Intelligence'

- ▶ introduces what is usually called the 'Turing Test' to replace the question 'Can machines think?'
- ▶ Turing described 'The Imitation Game': a man (A), a woman (B) and an interrogator (C) who must decide whether X is A and Y is B or vice versa.
- ▶ questions put to both A and B: A is trying to persuade C to make a mistake, B is trying to help C.
- ▶ If we instead have A is machine, B is human, how often will C get the identification wrong (after 5 minutes)?

Intelligence as ungrounded imitation?

- ▶ Turing described an abstract test (avoiding the complications of robotics, vision etc).
- ▶ But communication is central.
- ▶ Deception is key to the test: computer 'pretends' to be human.
- ▶ Sloman (e.g., p606–610 Cooper and van Leeuwen (eds), 2013) argues that Turing did NOT propose this as a test for intelligence.
- ▶ Searle 'Chinese Room': discussion of consciousness, criticism of Strong AI.

Lexical meaning: what doesn't work

- ▶ meaning of tomato is tomato'?
- ▶ meaning postulates (lecture notes! What is a chair?)
- ▶ dictionary definition
  tomato: mildly acid red or yellow pulpy fruit eaten as a vegetable
  good dictionary definition allows reader with some familiarity with a concept to identify it

Lexical meaning: unanswered questions

- ▶ how far does distributional semantics get us?
- ▶ grounding often claimed for systems combining vision and language: is this enough?
- ▶ are virtual worlds a possible basis for grounding?
- ▶ or do we really need robots?