Lexical Semantics

Jonathan May

October 10, 2023

1 Semantics

This is potentially a huge topic – note that it takes up three chapters of Eisenstein. Semantics is the study of how to understand the meaning behind language. The question of what meaning even is gets a bit philosophical and this is part of why the topic is potentially large. We can think of semantics, or rather semantic analysis (per Eisenstein's description) as converting natural language into a *meaning representation* that connects to knowledge about the world. Further, each thing that is known should have a single representation such that if the representation changes, the meaning changes.

What is the value of this kind of concrete semantic analysis? We want to do things like talk to robots, ensure translation is faithful to meaning, remove or add bias, author style, etc. Checking the underlying meanings haven't changed or that they reflect what is intended seems intuitively like a good idea. Other ideas? If we can do end-to-end tasks, is semantic analysis (or syntactic analysis) still meaningful?

One thing I won't discuss is Logical semantics – the interpretation of natural language as a set of logical formulas including negation, conjunction, disjunction, implication, associativity, etc. You can read more about this in chapter 12 of Eisenstein.

Here are some other parts of meaning.

2 Semantic Similarity

To begin with, instead of analyzing each word in myriad dimensions of potential meaning, let's simply discuss the relative relationship of words to each other. If two words are interchangeable we can say they are 'synonyms.' Why is it useful to determine these? Consider a question answering task. We want to answer this:

What is a good way to remove wine stains?

We can use a few rules and search a big corpus for sentences starting 'A good way to remove wine stains is.' But we can do better. We'd like to match any of these as well:

Salt is a great way to eliminate wine stains

How to get rid of wine stains

How to get red wine out of clothes

Oxalic acid is infallible in removing iron-rust and ink stains

Knowing that 'remove' is synonymous or at least similar to 'eliminate', 'get rid of', 'get out of', 'removing' and that 'good' is similar to 'great' would help in approximate matching.

You could use this to find movie recommendations; find movie scripts with similar words (or phrases, sentences, paragraphs) to scripts of movies you like.

We're pretty good at doing this. You probably don't need to be told which of these pairs are similar and which are not similar:

bank-money	doctor-beer
apple-fruit	painting-January
tree-forest	money-river
bank-river	apple-penguin
pen-paper	nurse-fruit
run-walk	pen-river
mistake-error	clown-tramway
car-wheel	car-algebra

It turns out humans ranking similarity on a scale from 0 to 4 do so with correlation of 0.9! that's pretty good!

What makes words similar?

- Meaning: (wait isn't all semantics meaning?) e.g. 'want' vs 'desire'
- World knowledge: things go together (pen-ink, dog-cat)
- psychology: we think of the concepts together (death-taxes)

2.1 Brief bit on linguistic terms

These definitions are sometimes helpful. You are probably familiar with some if not all of them from grade school.

- homonym: two words with same form but unrelated, distinct meanings.
  - homograph: bank (finance) vs bank (slope); bat (wood stick) vs bat (animal)
  - homophone: write vs right, piece vs peace
- polysemy: having more than one related meaning. bank (financial institution vs physical building)
- metonymy: one thing standing in for another ('I love Jane Austen'('s writing)), which can lead to introduction of senses e.g. 'school' to mean the two bank senses
- synonym: two words with same meaning
- antonym: two words with opposite meaning

2.2 Evaluation

Some ways to evaluate:

- Given a word and a choice of other words, find the one that is closest in meaning.
  - **accidental**: wheedle, ferment, inadvertent, abominate
  - **imprison**: incarcerate, writhe, meander, inhibit
  - WS353: a dataset of similarity scores for 353 English word pairs. Can be used to automatically create these tests.
- Malapropism test: find the word in the sentence that is most likely wrong
  - Jack withdrew money from the ATM next to the band.
  - Can be created by randomly replacing words from a lexicon

2.3 Hand-Built Resources

We cared about annotating semantic relationships between words, so much so that considerable effort was put into hand-crafting ontologies. For lexical semantics, the most famous (other than roget's thesaurus) was **WordNet**. It has 118k English nouns, 11.5k verbs, 22.5k adjectives, 5k adverbs.

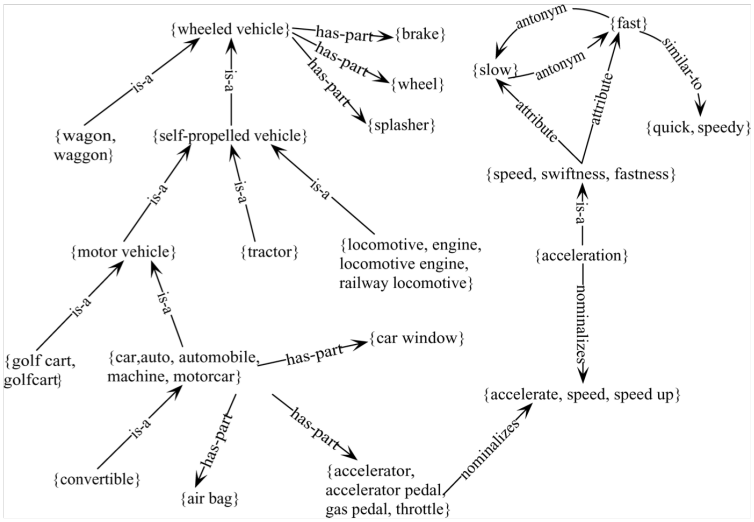
Lemmas (base word forms) have one or more senses (distinct meaning units), each with examples explaining the distinct meaning. Meaning units are known as *synsets*. A synset has a definition and often some examples. A lemma has multiple senses. Each sense has one synset. But different senses of different words can share the same synset.

E.g. chump-1 has the synset defined as "a person who is gullible and easy to take advantage of". This synset is shared by fool-2, gull-1, mark-9, patsy-1, fall guy-1, sucker-1, soft touch-1, mug-2.

Senses are structured in hypernym trees; a hypernym is more inclusive, a hyponym less inclusive. *car* is a hyponym of *vehicle*. Alternately this is an 'is-a' hierarchy (car is a vehicle). Alternately it's an entailment hierarchy (being a car entails being a vehicle)

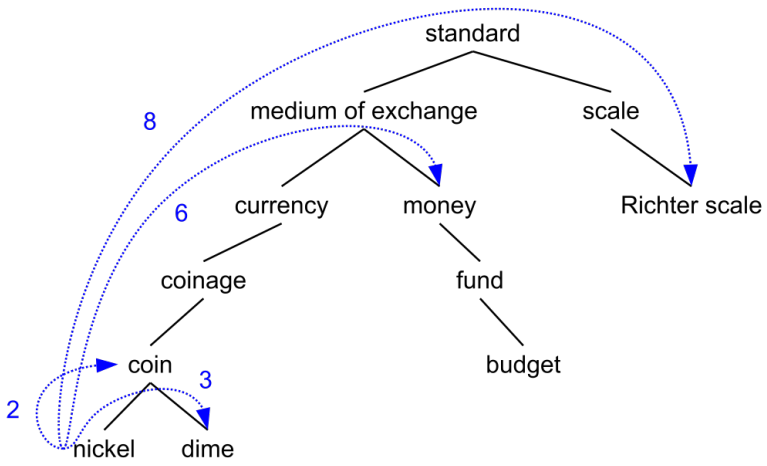
Wordnet also encodes *meronymy* (but less so); this is a part-whole relation; *leg* is a meronym of *chair*. Alternately, *chair* is a *holonym* of *leg*.

There are verb relations too but they're more incomplete.



Wordnet is incomplete and inconsistent; some parts are very dense, others have gaping holes. It is also increasingly out-of-date (e.g. *television* has meronym *kinescope*—these haven't been part of televisions for years!). Nevertheless it can be a very precise (subject to datedness) repository of info and can be used to determine word similarity with some quite simple algorithms over synset trees.

E.g. Path length == number of arcs you walk to get between words. A simple normalized refinement is *simpath* = 1/pathlen. It's a number from 0 to 1 and the closer words are, the higher the value is.



2.4 Distributional Methods

Normally I'd now talk about PPMI and then Word2Vec. But we already did that!

3 Semantic Roles and Representations

Rather than simply saying how 'close' two words are or giving each word a unique meaning based on its context (e.g. in a transformer or LSTM representation) we can analyze words

by the roles they play together in explaining the meaning of a concept. Here's an example sentence (taken from Nathan Schneider):

Mary loaded the truck with hay at the depot on Friday. (1)

We can syntactically analyze this but the three PPs aren't really distinguished. Note that some of them are moveable with the same meaning:

Mary loaded the truck with hay on Friday at the depot. (2)

but some moves change the meaning

Mary loaded the truck at the depot with hay on Friday. (3)

and some moves become (semi-)ungrammatical

?Mary loaded the truck at the depot on Friday with hay. (4)

Clearly more is going on than simple syntax.

3.1 PropBank

The Proposition Bank (PropBank[1]) is an ontology of (mostly) *predicates* (verbs or verb phrases divided by senses) and the arguments they bear. Predicates and senses are similar to those seen in WordNet: Here are senses for 'load'

- load-01: cause to be burdened ("UPS drivers aren't permitted to load their own vehicles at the depot"
- load_up-02: (phrasal) cause to be burdened ("KKR loaded up the cable and television company with debt")
- load-03: fix, set up to cheat ("If you wish to be able to repeat the trick right away, you need to have loaded the deck with 10 pre-arranged cards on top.")
- load-04: transfer data from one place to another ("The site loads soo slowly on my phone!")

But that's not all! Along with each sense there is a set of *roles*, which are words/phrases that are important semantic modifiers of the predicate. You can think of the predicate as like a function and the roles as arguments...in fact we call them that!

We have (numbered) roles that are specific to predicates, since not all roles make sense for all predicates. For example 'load-01' has ARG0 (loader), ARG1 ('beast' of burden), ARG2 (cargo), and ARG3 (instrument). Whereas 'load-03' has ARG0 (cheater), ARG1 (thing loaded), and ARG2 (with what). Note that these do have some rough correspondence, but it varies: 'put-01' has ARG1 (thing put) which is like the ARG2 of 'load-01'.

[1] `https://propbank.github.io/`

There are also non-numbered roles that can go with all (or most) predicates: TMP (when?), LOC (where?), DIR (where to/from?), MNR (how?), etc.

So we have a new task! Given a sentence, find the predicates, classify them into their sense, then identify their labeled arguments! In sentence 2, 'loaded' is load-01, with ARG0 'Mary', ARG1 'the truck', ARG2 'hay', LOC 'the depot', and TMP 'Friday.'

There are a few SRL tasks. Ontonotes, for example, has Treebank and Propbank annotations for English, Arabic, and Chinese (1.4m+ words of English) and is actively improved (88.1 f-measure new as of 2022).[2]

Each word/phrase type is considered a 'frame' and there are more than 7500 English PropBank frames (and it's still actively developed). It exists for some other languages too (but not very many).

3.2 FrameNet

A 'competitor' is FrameNet,[3] which tries to be 'richer' and as a result has less coverage. The idea of FrameNet is that the meaning of a text forms a 'scene' and so different texts with the same scene should have the same analysis. Consider:

Mary filled the vehicle with hay at the station on Friday. (5)

As the meaning of sentences 2 and 5 are the same, they should have the same analysis. FrameNet defines a frame called 'Filling'.[4] It provides a definition ('These are words relating to filling containers and covering areas with some thing, things or substance') and has *frame elements*, the analogue of roles, but with interdependence and sometimes extra explanation:

- Agent – the actor who instigates the filling.
- Cause – An event which brings about the filling of the Goal.
- Goal – The Goal is the area or container being filled. Goal is generally the NP Object in this frame.
- Theme – The Theme is the physical object or substance which changes location.

Those are the core elements, but there are other 'non-core' roles that also reference the other roles. For example, Purpose is 'The purpose for which the Agent fills the Goal.'

There are more than 1,200 frames. Frame semantic parsing of FrameNet seems less popular than PropBank but there is still active development (81.7 F1 as of 2022 but I didn't read this paper closely[5]).

[2] `https://paperswithcode.com/task/semantic-role-labeling`

[3] `https://framenet.icsi.berkeley.edu/`

[4] `https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Filling`

[5] `https://arxiv.org/pdf/2206.09158.pdf`

3.3 Abstract Meaning Representation

An attempt to perform whole-sentence semantic analysis (not just predicates and arguments), AMR[6] incorporated PropBank roles and senses but also entity annotation and coreference, negation, questions, etc. Because of coreference, the annotation can form a graph.

This will ultimately accelerate the speed of desertification in sub-Saharan African countries and other areas of the world. (6)

```
(a / accelerate-01
      :ARG0 (t / this)
      :ARG1 (s / speed-01
            :ARG0 t
            :ARG1 (d / desertification
                  :location (a2 / and
                        :op1 (c / country
                              :location (w2 / world-region :wiki "Sub-Saharan_Africa"
                                                           :name
                                                     (n / name :op1 "Sub-Saharan"
                        :op2 (a3 / area
                              :part-of (w / world)
                              :mod (o / other)))))
      :time (u / ultimate))
```

More conventional:

The soldier did not want to die. (7)

```
(w / want-01
    :arg0 (s / soldier)
    :arg1 (d / die-01
          :arg0 s)
    :polarity -)
```

The AMR corpus consists of about 60k annotated sentences from a variety of genres. Parsing English into AMR and generating English from AMR are active tasks. 86.7 smatch (an F1 of 'mappable' graph elements) for parsing[7] and 49.8 BLEU on generation[8].

[6] https://amr.isi.edu/

[7] https://paperswithcode.com/sota/amr-parsing-on-ldc2017t10

[8] https://paperswithcode.com/sota/amr-to-text-generation-on-ldc2017t10

3.4 Crowd-Sourced Repositories

WordNet, Cyc (which I haven't said much about), PropBank, FrameNet, and AMR were all careful, commissioned work. As such they tend to grow slowly and may reflect more of an ivory tower mentality than a bottom up understanding of meaning. The following resources emerged in a semi-bottom-up way.

3.4.1 DBPedia

Some folks based in Germany starting in 2007 decided to expose wikipedia info boxes which have structured text. The info boxes constitute over four million entities (e.g. place, music album, organization, species) and labeled information within them constitutes arcs to other entities. For example, from the info box below we can say that the album 'Red' is on the label 'Big Machine'; this can be viewed at `https://dbpedia.org/page/Red_(Taylor_Swift_album)` along with all the other properties of that album.

Currently DBpedia has over 21 billion triples in it, which are released monthly, and releases a downloadable 'snapshot' of 1 billion triples of mostly english plus multilingual abstracts in a number of other languages, as well as links to other data sets in the Linked Open Data Cloud – data sets like the CIA World Fact Book, Project Gutenberg, GeoNames, the US Census, and lots of others. There is lots of noise; for example there may be redundant fields like 'placeofbirth' and 'birthplace', or information may be wrong or missing. The developers provide endpoints to access this data, attempts to clean up redundant concepts,

and other support. As of 2023 this is under heavy active development and is well supported.

3.4.2 ConceptNet

From 1999 to 2016 the Open Mind Common Sense website anyone could go to a web site and contribute common sense knowledge about the world. You would be given a prompt like 'A hammer is for __' and fill in the blank. Or a simple story would be started like 'Bob had a cold. Bob went to the doctor.' and you would be asked to write some other consistent simple sentences. Or a photo would be shown and people would describe the photo. It's kind of shocking to me that this worked without having to pay people and didn't get overrun with spam, but over 17,000 people provided hundreds of thousands of entries. The data source has since been expanded to include portions of DBPedia and Wiktionary. We can play around with it at `https://conceptnet.io/`.