Explicit world-knowledge and distributional semantic representations

ESSLLI 2017 Day 5: Modeling the distinctions

Asad Sayeed and Alessandra Zarcone

University of Gothenburg, Saarland University

So now we get into the speculative part of our course.

Part 1: unexplored experimental avenues

Plausibility

We left the last lecture with the following:

- Amsel et al. found that "perceptuomotor" and "event-related" anomalies produced N400 in different parts of the brain.
- However, event-based and perceptuomotor-based anomalies are hard to distinguish when plausibility-ratings are matched.

So maybe plausibility is not a "real" thing, only mismatches matter.

Doch.

Event-relatedness vs. perceptuomotoricity – too fine-grained a distinction.

- What we really want to measure: execution of higher-order affordance knowledge (= plausibility)

(1) Bob cut a cake

a. with a knife. (typical/frequent, possible)
b. with a hammer. (distributionally similar to *knife*, impossible)
c. with floss. (atypical/dissimilar/infrequent, possible)
d. with a towel. (atypical/dissimilar/infrequent, impossible)

(A note on possibility and plausibility.)

The distinction is nitpicky, but here's one definition for this purpose:

Possibility

The absolute ability of the listener to execute affordance knowledge in a way that "converges" on an imagined event.

Plausibility

The relative effort in executing affordance knowledge in the context of a possible event.

Plausibility is more interesting as it focuses on the "mechanics", but we can use possibility as a lever.

So back to our paradigm.

(2) Bob cut a cake

a. with a knife. (typical/frequent, possible)
b. with a hammer. (distributionally similar to *knife*, impossible)
c. with floss. (atypical/dissimilar/infrequent, possible)
d. with a towel. (atypical/dissimilar/infrequent, impossible)

The problem with this: extremely difficult to come up with EEG or even eyetrack-worthy balanced data sets.
Even Amsel et al. put in a lot of effort to come up with their stimuli – over various senses/event relations.

Just use thematic fit!

Instead of trying to carefully norm a balanced possibility data set:

- We already have testing data for our distributional models: thematic fit ratings!
- Likert scale 1-7 averaged over 20-ish humans of event triples: (verb, role, filler) = score.
- The problem is:
  - Some of them are normed for "commonness": "How common is it for ..."
  - The ones built by Greenberg, Demberg, and Sayeed were slightly more neutral: "An X is something that is Y'd" – eliminate "first-order" affordance knowledge bias.

Just use thematic fit!

Proposal: collect ratings for an instrument data set.

(3) Rate from 1-7:

a. How common is it to use a knife to cut something? (probability-biased)
b. A knife is something that you use to cut. ("equipoise")
c. Can a knife be used to cut? (possibility-biased)

Then use these ratings for correlation studies **both** across computational models and psycholinguistic measures.
Can use McRae instrument data via crowdsourcing for starters, 200-300 ratings.

How do we use these ratings?

For psycholinguistic measures: possibly combine with highly-rated objects?

(4) a. Bob cut a cake with a knife
b. Bob cut a cake with a string
c. Bob cut a cake with a committee
d. Bob cut a budget with a knife
e. Bob cut a budget with a string
f. . . .

Or even just try to do it without an object role.

Cake with hammer

How do we use these ratings?

Also, to evaluate computational models:

- Ratings that contrast distributional and “plausibilistic” intuitions by humans are in themselves valuable.
- What are statistical models of semantics actually capturing?
  - Difference between model correlations with probability-biased and possibility-biased ratings = influence of non-distributional knowledge?
  - Does improvement on one reflect improvement on another, or are we just building smarter and smarter “abstract parrots”?

Formal representation

But how do we capture those "plausibilistic" or affordance-based intuitions?

- Generative Lexicon qualia structure seem to be a good place to put them.
- Problem is, which quale? (Constitutive, Formal, Telic, Agentive)
- Knives are for cutting (Telic Quale) but...
  - Shape affords actions that don't fit GL qualia structures.
  - E.g., hold them in a hand, throw them, insert them in knife block, etc.

Part 2: multimodal approaches to feature extraction

If we need real-world data, from where can we get it?

Start from the most obvious: image data.

- So yes, image data is somewhat unambitious.
  - Static, decontextualized.
  - Are there any ways in which the distribution of this data may be skewed?
- On the other hand: it's nearly as abundant online as textual data.

There was always a motivation for doing this: grounding AI.

- No matter how formal the semantics you use, you're going to need to connect it to the real world.

How to use image data?

- Text-image link: what consitutes an image linked to a text? Same document? Tweets? Human selection?
- Actually using the image:
  - Learn on labelled data – expensive, but you can get lots of image labels off the internet.
  - Learn on pixels – ideal, but computational intensive and with data sparsity issues.

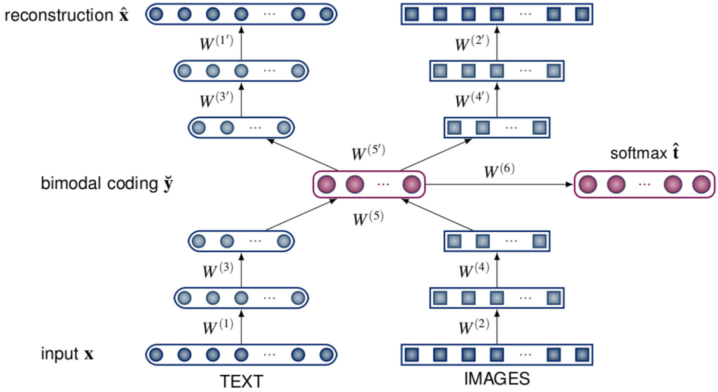
Learning on labelled data

Multimodal Deep Learning. Silberer and Lapata [2014]:

- Use feature-attribute norms from McRae et al. [2015].
  - 541 nouns with human-labelled attributes.
  - Literally the whole “define chair’ exercise, basically.
- 700K images from ImageNet [Deng et al., 2009] labelled with 636 visual attributes.
- 2362 textual attributes extracted from Wikipedia.
- “Stacked bimodal autoencoder” to learn both representations.
- Long story short: autoencoder has highest correlation with human ratings w.r.t. previous work.

Learning on labelled data

Stacked bimodal autoencoder from Silberer and Lapata [2014]:



Learning from pixels

A word2vec-style approach from Lazaridou et al. [2015]:

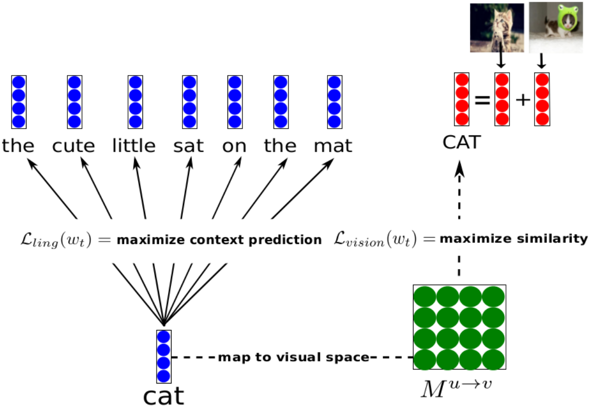
- Multimodal skip-gram model based on Mikolov et al. [2013a].
- Eval data: approx 12K semantic relatedness judgements (e.g. pickles are related to hamburgers).
- Text corpus: 800M-token Wikipedia dump.
- Image corpus: 5100 visual representations derived from ImageNet.

Long story short:

- Performs on human judgements within range or better than competing systems.
- Outperforms basic skip-gram model in image labelling and retrieval tasks.

Learning from pixels

Lazaridou et al. [2015] model sketch:



Part 3: concluding remarks

Types of knowledge

Dividing up the knowledge problem:

Implicit world-knowledge

Latent knowledge about the world that can be induced from indirect information sources (e.g. distributional characteristics of language).

Explicit world-knowledge

Knowledge about the world that is coded explicitly, deduced formally, innate, learned by being told, etc.

- Implicit world knowledge – somehow related to the "experiential" part of extensional meaning?
- Explicit world knowledge – somehow related to the "cognitive" part of intensional meaning?

How far does distributional semantics get us?

Quite far:

- Can use it to characterize similarity – that's what all the word embeddings craze is ultimately about.
  - Similarities are great for many, many tasks.
- Generalized event knowledge – works quite well, provided we have:
  - a semantic space that is sufficiently well-structured/informative.
  - some procedure for exploiting the structure of the space.
  - some finer classification of events and entities.

That's explicit knowledge, but not a lot of explicit knowledge – fair enough.

How far does distributional semantics get us?

Not far enough:

- Similarity, generalized event knowledge not enough to help us with huge domains of interaction.
- Affordances, plausibility:
  - People can "simulate" the "misuse" of unexpected objects.
  - How would you replicate/model this behaviour "distributionally"?
  - Very preliminary approach: augmentation with multimodal distributional data.

Maybe formal semantics can help?

Formalisms like Generative Lexicon, compositional approaches etc.

- Ways to represent script knowledge, expectations.
- Central problem: "turtles all the way down".
  - What is the empirical basis to justify the primitives of the formalism?
- Nevertheless, it seems like we need some kind of formal structure to characterize state change.

The problem is very far from "solved," but there are lots of opportunities for research.

Questions and plenary discussion (if there's time) Otherwise: THANKS