Advanced Machine Learning, Lecture 10 (week 11)

András Kornai

BME 2020 Nov 19

Improving static word vectors by disambiguation

- Why is this important when we already have the dynamic vectors?
- There is reason to believe that whatever is stored in memory is static, with higher representations computed from these dynamically
- We want to understand what is going on (BlackBox AI, XAI)
- Experimentation has started with DilBERT, a static but multi-sense system
- The aim is to measure the delta between dynamic and true multi-sense
- BERT and similar systems start with a static embedding at the lowest layer, but we don't normally have the means of swapping this out and retrain the entire system

Refresher: word vectors and polytopes

- The word vector partition the space into polytopes
- We think of each word labeling not just its vector but the entire polytope surrounding it
- Consider *bark*. In a multi-sense system this is not just one vector but two, corresponding to $bark_1$ 'fakéreg' and $bark_2$ 'ugat(ás)'
- The polytopes surrounding these vectors are not even close to one another: lexicographers speak of *homonymy* (shared form)
- There are many cases, where separate vectors can be assumed, e.g. *prison* can mean either 'building to keep convicts in' or 'punishment by inprisonment'. There are many contents in which it is not clear which sense is meant: *He was sentenced to five years prison*. Lexicographers call this *polysemy*. The polytopes are adjacent/overlapping

How to distinguish the cases?

- In case of homonymy there will be many hyperplanes that separate the two polytopes: e.g. $bark_2 \subset$ sound, $bark_1 \not\subset$ sound
- Separating half-space is not unique: e.g. ±PHYSOBJ also separates the two, as does `dog make` and many other predicates
- Using HuBERT, we can lok at many contextual instances, and see if they can be separated by large margin
- Early attempts at multi-sense embeddings: Reisinger and Mooney 2010; Huang et al 2012 – k senses for every word
- Neelakantan et al 2014; Li and Jurafsky 2015; Bartunov et al 2016 – k can be different for diferent words
- These don't match actual word-sense distribution (Borbély et al 2016)
- Wins from the better methods are comparable to having embeddings with more dimension (Li and Jurafsky 2015)

Improving embeddings by including further data sources: dictionaries

- Dictionaries can be used to generate embeddings entirely on their own: reduce dictionary definitions, treat as edge-weighted graph (Ács et al 2017) or by more sophisticated methods (Kornai in prep)
- We can use word vectors generated in different languages

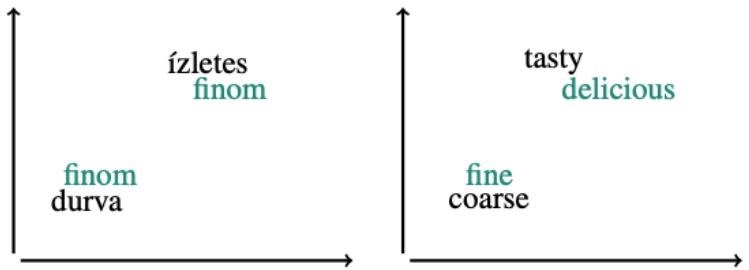

Figure 1: Linear translation of word senses. The Hungarian word *finom* is ambiguous between 'fine' and 'delicious'.

(Makrai and Lipp 2017)

- We can add dictionary similarity to the training objective (Tissier et al 2017)

Improving embeddings by including further data sources: knowledge graphs

- Celikyilmaz et al 2015
- Long et al 2016
- Ahn et al 2017 *Kanye West, a famous UNKNOWN and the husband of UNKNOWN, released his latest album UNKNOWN in UNKNOWN*
- Dinghra et al 2017