Vector Semantics

Tim Lawson

October 20, 2023

The distributional hypothesis

"For a *large* class of cases of the employment of the word "meaning" — though not for *all* — this word can be explained in this way: the meaning of a word is its use in the language."[1]

[1] Wittgenstein 2010, p. 25.

The distributional hypothesis

"...we will often find interesting distributional relations, relations which tell us something about the occurrence of elements and which correlate with some aspect of meaning. In certain important cases it will even prove possible to state certain aspects of meaning as functions of measurable distributional relations."[2]

[2] Harris 1954, p. 156.

The distributional hypothesis

"In these examples, the word *ass* is in familiar and habitual company, commonly collocated with *you silly*—, *he is a silly*—, *don't be such an*—. You shall know a word by the company it keeps!"[3]

[3] Firth 1957, p. 11.

Distributional (vector-space) semantics

- What aspects of meaning can we talk about?
- What "distributional relations" can we use?
- How do choices of relations affect the outcomes?
- What can we do with a vector? With multiple vectors?

Word embeddings

- Term-document and term-term vectors
  - Many-dimensional ($d =$ number of terms $\times$ number of documents)
  - Sparse (mostly zero)
- Word embeddings
  - Lower-dimensional ($d \approx 50 - 1000$)
  - Dense (mostly non-zero)
- "Dense vectors work better in every NLP task"[4]

[4] Jurafsky and Martin 2023, p. 119.

Skip-gram with negative sampling

- Skip-gram: “maximize[s] classification of a word based on another word in the same sentence”[5]
- Negative sampling: “distinguish[es] the target word w_0 from draws from the noise distribution $P_n(w)$ ... where there are k negative samples for each data sample”[6]

[5] Mikolov, Chen, et al. 2013, p. 4.

[6] Mikolov, Sutskever, et al. 2013, p. 4.

Skip-gram with negative sampling

Feature engineering

- Find unigram counts and weights
- Find n-grams where $n - 1$ is the window size
- Find positive examples (target-context word pairs)

Skip-gram with negative sampling

Training

- Initialize target- and context-word matrices
- Iterate until stopping criterion is met
  - Update the positive context-word vector
  - Find negative examples by sampling from the vocabulary
  - Update the negative context- and target-word vectors

Semantic similarity

- First-order co-occurrence (syntagmatic association)
  - “I *wrote* a *book*.”
- Second-order co-occurrence (paradigmatic association)
  - “I *wrote/read* a book.”

Semantic similarity

- Syno- and antonyms
  - “I took the *bus/coach*.”
  - “The matrix is *sparse/dense*.”
- Hyper- and hyponyms
  - “I had a *coffee/espresso*.”
- Metonyms
  - “*Number 10/The government* said that...”

Visualizing embeddings

- Nearest neighbours
- Clustering
- Dimensionality reduction

Bibliography

- Firth, J. R. (1957). “A Synopsis of Linguistic Theory, 1930–1955”. In: *Studies in Linguistic Analysis*, pp. 1–31.
- Harris, Zellig S. (1954). “Distributional Structure”. In: *WORD* 10.2-3, pp. 146–162.
- Jurafsky, Dan and James H. Martin (2023). *Speech and Language Processing*.
- Mikolov, Tomas, Kai Chen, et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: `1301.3781 [cs]`.
- Mikolov, Tomas, Ilya Sutskever, et al. (2013). “Distributed Representations of Words and Phrases and Their Compositionality”. In: *Advances in Neural Information Processing Systems*. Ed. by C. J. Burges et al. Vol. 26. Curran Associates, Inc.
- Wittgenstein, Ludwig (2010). *Philosophical Investigations*. John Wiley & Sons.