Tackling fairness, change and polysemy in word embeddings

Felipe Bravo-Marquez

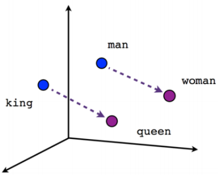
Department of Computer Science, University of Chile
National Center for Artificial Intelligence Research
Millenium Institute Foundational Research on Data

December 3, 2021

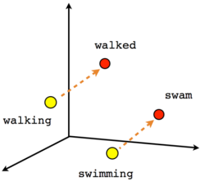
Word Embeddings

- The first step in computationally working with written language is to represent words as mathematical objects we can operate with.
- Representing words as numeric vectors a.k.a **embeddings** is a standard practice in **Natural Language Processing** (NLP).
- Word embeddings are a mapping of discrete symbols (i.e., words) to continuous vectors.
- Distance between vectors can be equated to distance between words.
- This makes easier to generalize the behavior from one word to another.
- Word embeddings have become a core component of natural language processing (NLP) downstream systems (e.g., sentiment analysis, machine translation, question answering).

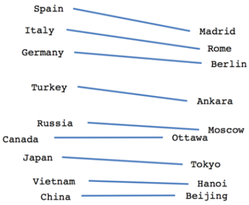
Word Embeddings



Male-Female



Verb tense



Country-Capital

Word embeddings can encode semantic and syntatic relationships between words.

Distributional Hypothesis

- The construction of word embeddings from document corpora is based on the **Distributional Hypothesis** [Harris, 1954]:

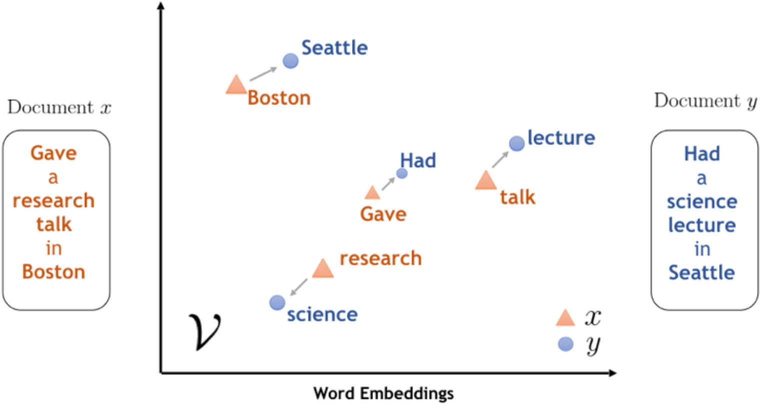
  *Words occurring in the same **contexts** tend to have similar meanings.*

- Or equivalently:

  *A word is characterized by the **company** it keeps.*

- The idea is to map words occuring in similar contexts to similar vectors.

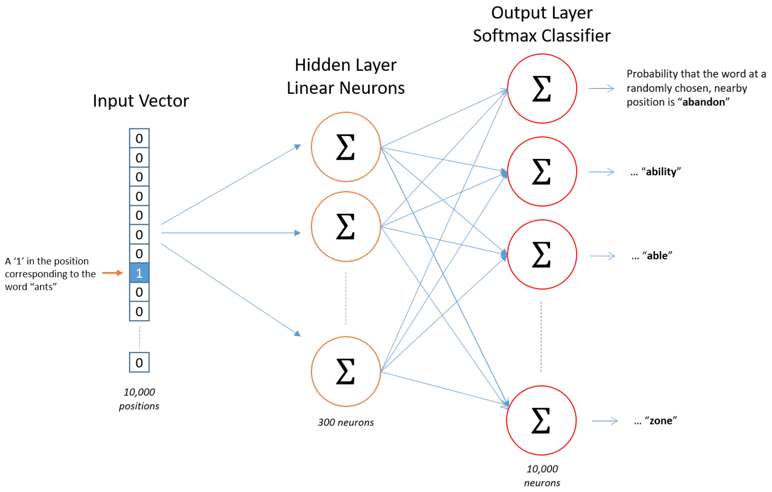
Word Embeddings



Word Embeddings Algorithms

- Word embeddings are build by training neural networks architectures on document corpora (e.g., books, papers, Wikipedia, tweets, the Web).
- These arquitectures formulate a predictive task (e.g., predict a missing word withing a contex window) in which word embeddings naturally arise from the network's parameters after training.
- Most popular models are:
  - Skip-gram negative sampling [Mikolov et al., 2013]
  - Continuous bag-of-words [Mikolov et al., 2013]
  - Glove [Pennington et al., 2014]
  - FastText [Bojanowski et al., 2016].

Skip-gram Model



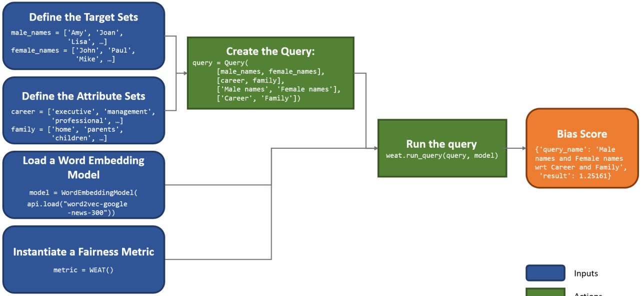
[0] Picture taken from: `http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/`

Limitations of Word Embeddings

- However they suffer from three major limitations:
  1. **Fairness**: they are prone to inherit stereotypical social biases from the corpus they were built on.
  2. **Change**: they are static. Thus they ignore words not observed during training and are unable to capture semantic drifts.
  3. **Polysemy**: they fail to capture the polysemous nature of many words (e.g., apple:company, apple:fruit), conflating their multiple senses into a single point.
- In this talk we will present our research addressing these three problems.

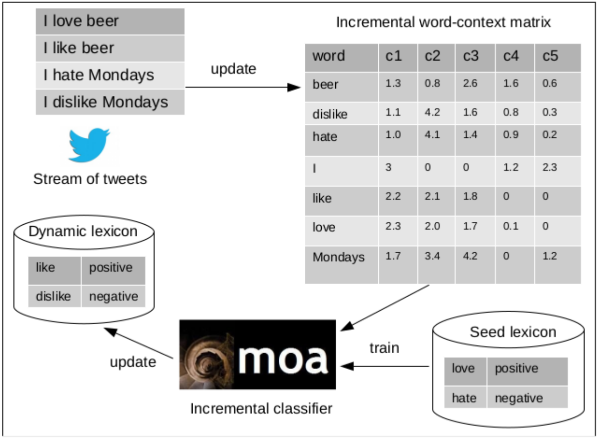
WEFE: The Word Embeddings Fairness Evaluation Framework

- The Word Embeddings Fairness Evaluation (WEFE) is a framework for measuring and mitigating bias in word embeddings (e.g. man is to programmer as woman is to housewife). [Badilla et al., 2020].



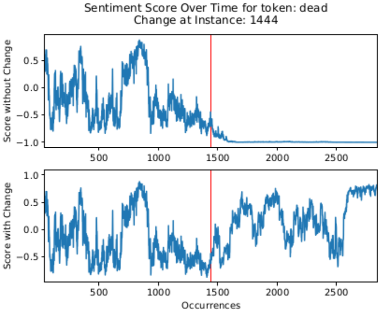
Incremental Word Vectors

- An algorithm capable of continuously learning word vectors and thus understanding how the meaning evolves over time (e.g., monitoring the word "estallido" in social networks during the Chilean social unrest). [Bravo-Marquez et al., 2021].



Incremental Word Vectors

- We simulate sentiment change by randomly picking some words and swapping their context with the context of words exhibiting the opposite sentiment.

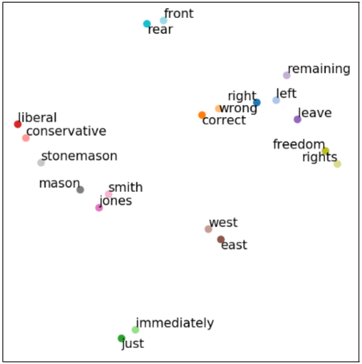


1. (a) dead

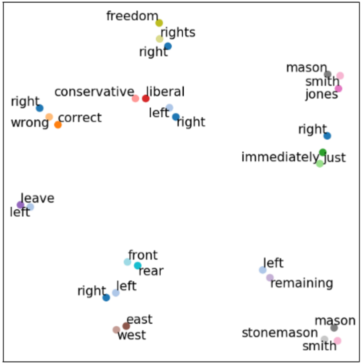
- Our approach allows for successfully tracking of the sentiment of words over time even when drastic change is induced.

PolyLM: a polysemous language model

- A language model capable of automatically learning multiple meanings of a word (e.g. apple:apple, apple:company) [Ansell et al., 2021].

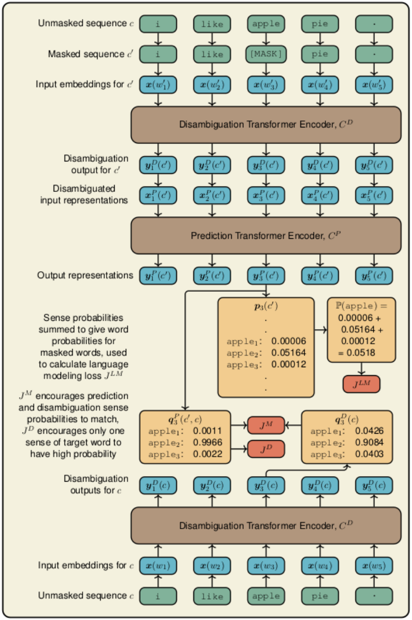


(a) Word embeddings



(b) Sense embeddings

PolyLM: a polysemous language model



Questions?

Thanks for your Attention!

References I

Ansell, A., Bravo-Marquez, F., and Pfahringer, B. (2021).
Polylm: Learning about polysemy through language modeling.
In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 563–574.

Badilla, P., Bravo-Marquez, F., and Pérez, J. (2020).
Wefe: The word embeddings fairness evaluation framework.
In *IJCAI*, pages 430–436.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016).
Enriching word vectors with subword information.
arXiv preprint arXiv:1607.04606.

Bravo-Marquez, F., Khanchandani, A., and Pfahringer, B. (2021).
Incremental word vectors for time-evolving sentiment lexicon induction.
Cognitive Computation, pages 1–17.

Harris, Z. (1954).
Distributional structure.
Word, 10(23):146–162.

References II

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Pennington, J., Socher, R., and Manning, C. D. (2014).
Glove: Global vectors for word representation.
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.