Tackling fairness, change and polysemy in word embeddings

Felipe Bravo-Marquez

Department of Computer Science, University of Chile
Millenium Institute Foundational Research on Data

November 22, 2021

dcc
CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Millennium Institute
Foundational
Research on Data

RELELA
Representations for
Learning and Language

Word Vectors

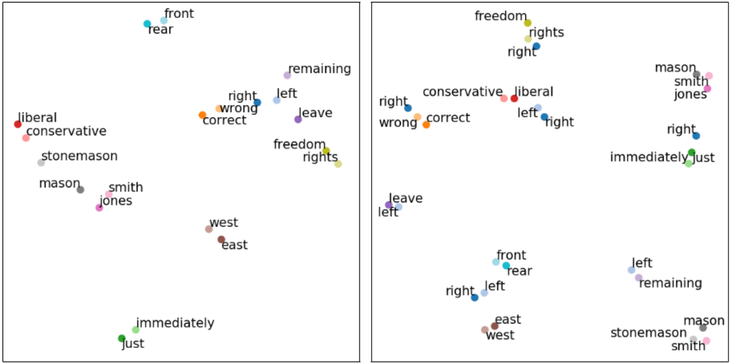
- A major component in neural networks for language is the use of an embedding layer.
- A mapping of discrete symbols to continuous vectors.
- When embedding words, they transform from being isolated distinct symbols into mathematical objects that can be operated on.
- Distance between vectors can be equated to distance between words.
- This makes easier to generalize the behavior from one word to another.

Distributional Vectors

- **Distributional Hypothesis**: words occurring in the same **contexts** tend to have similar meanings.
- Or equivalently: “a word is characterized by the **company** it keeps”.
- In this talk we summarize our research addressing three limitations of static word embeddings: 1) fairness, 2) semantic change, and 3) polysemy.

PolyLM: a polysemous language model

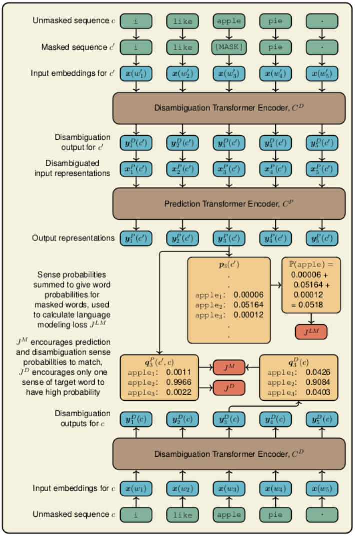
- A language model capable of automatically learning multiple meanings of a word (e.g. apple:apple, apple:company) [Ansell et al., 2021].



(a) Word embeddings

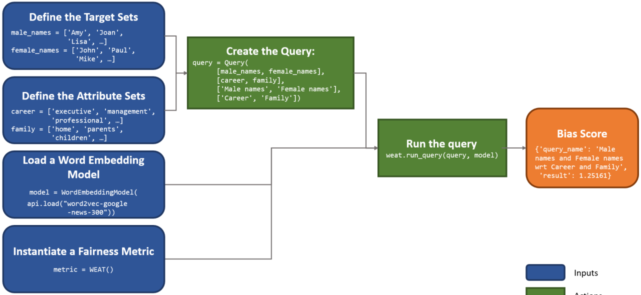
(b) Sense embeddings

PolyLM: a polysemous language model



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].



Define the Target Sets

```
male_names = ['Amy', 'Joan',
              'Lisa', …]
female_names = ['John', 'Paul',
                'Mike', …]
```

Define the Attribute Sets

```
career = ['executive', 'management',
          'professional', …]
family = ['home', 'parents',
          'children', …]
```

Load a Word Embedding Model

```
model = WordEmbeddingModel(
api.load("word2vec-google
          -news-300"))
```

Instantiate a Fairness Metric

```
metric = WEAT()
```

Create the Query:

```
query = Query(
      [male_names, female_names],
      [career, family],
      ['Male names', 'Female names'],
      ['Career', 'Family'])
```

Run the query

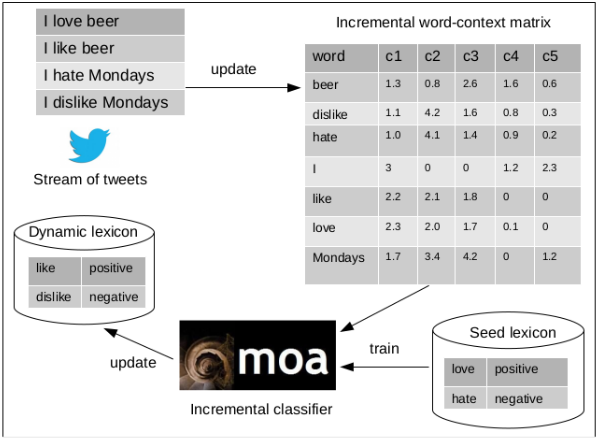
```
weat.run_query(query, model)
```

Bias Score

```
{'query_name': 'Male
names and Female names
wrt Career and Family',
 'result': 1.25161}
```

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].



I love beer
I like beer
I hate Mondays
I dislike Mondays

Stream of tweets

update

Incremental word-context matrix

word	c1	c2	c3	c4	c5
beer	1.3	0.8	2.6	1.6	0.6
dislike	1.1	4.2	1.6	0.8	0.3
hate	1.0	4.1	1.4	0.9	0.2
I	3	0	0	1.2	2.3
like	2.2	2.1	1.8	0	0
love	2.3	2.0	1.7	0.1	0
Mondays	1.7	3.4	4.2	0	1.2

Dynamic lexicon

like	positive
dislike	negative

update

Incremental classifier

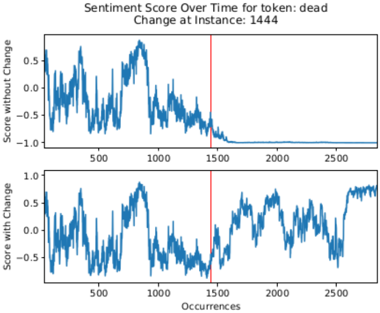
train

Seed lexicon

love	positive
hate	negative

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.

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.

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