Tracing and visualizing diachronic semantic change using contextualized embeddings

Software project, group 5

Averie (Ho Zoen) SO, NGO Van Duy, Scott TANKARD, Mathilde AGUIAR

Université de Lorraine

Friday 18th November, 2022

Outline

Introduction

Project formalisation

Working draft project goal proposal:
tracing and visualizing diachronic semantic change using contextualized embeddings (from m-BERT), with re-training on an array of multilingual time-segmented corpora
(With 1 model per time segment. Example: Model A: 1910 english + 1910 french. Model B: 1920 english + 1920 french.)
Tracing: putting in relation of multiple quantified (non-binary) measurements (of semantic change).
Underlying core-core part here is: get quantified measurements of semantic change from the model (m-bert).

Extra steps and components we may add

As first step: set-up with just 1 language, for 2 time periods. Later, add additional languages into the corpuses and re-train the models.
Bonuses:
+ Run experiments on multi-senses vs single-averaged sense (WITHOUT testing on different types of semantic change)
+ analyzing multiple languages in comparison to each other (e.g. evolution of Sir/Monsieur in eng/fr)
+ historical event contextualization (database...)
+ future semantic change prediction
Non-goals (explicitly excluded from project scope):
- exploring the multilinguality inside multilingual models
- doing multiple monolingual applications

HistWords

William L. Hamilton et al. 2018

Publication: *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change*

- Quantified semantic changes in word embeddings
- Experimented on 6 datasets in 4 languages: EN, ZH, DE, FR
- Historical embeddings are aligned with Proscrutes Regression
- Pairwise similarity is calculated using cos-sim

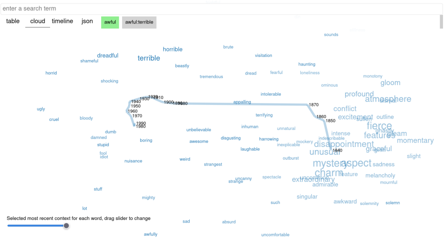


Figure: The illustration of AWFUL meaning shifts

BERT implementation

BERT implementation

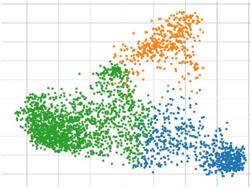
- Motivation
  - BERT is SOTA for word embeddings, but most literature in detecting semantic changes were before BERT existed (-2017)
  - contextualised word embeddings can potentially better capture the various senses of the same word
  - mBERT can deal with mapping embeddings across different languages
- Challenges and proposed solutions
  1. time bias in mBERT since it is pre-trained with more contemporary data $\rightarrow$ time-specific BERT
  2. tracing multiple word senses over time $\rightarrow$ clustering

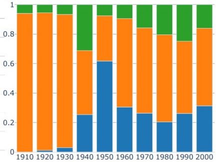
challenge 1: time-specific mBERT

- BERT is pre-trained on contemporary corpora (eg. Wikipedia) which often do not appropriately reflect language use in historical times. (Qiu and Xu, 2021)
- our attempt to alleviate this problem would be to further pre-train mBERT on time-specific corpora (unsupervised)
- given that we will use mBERT, we segment corpora from each language for each time period (eg. 1910 English + 1910 French, 1920 English + 1920 French..)
- then we train mBERT embeddings for each of these time periods multi-lingually

challenge 2: single vs multiple senses in mBERT

- single sense approach: various approaches in SemEval Task 2020 simply averaged across all contextualised embeddings of the same word
- multi-sense approach: clustering, eg. Giulianelli et al, 2020 (and others)


(a) PCA visualisation of the usage representations.



(b) Probability-based usage type distributions along time.

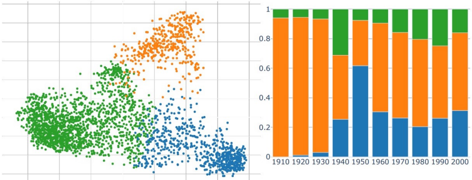
Figure: clustering and frequency distribution of senses for the word ATOM.

overall flowchart

1. obtain word embeddings from off-the-shelf mBERT, create a Usage matrix

period	sentence	mBERT word embeddings
1910	There's not an <u>atom</u> of dirt in her house	[3.3596…]
…		
2020	An <u>atom</u> is made up of protons, neutrons, and electrons.	[3.0123…]

2. clustering with k-means and silhouette score to select the optimal number of clusters and frequency distribution of word senses over time



(a) PCA visualisation of the usage representations.

(b) Probability-based usage type distributions along time.

overall flowchart (cont.)

3. obtain sense-tagged Usage matrix

period	sentence	mBERT word embeddings	sense
1910	There's not an **atom** of dirt in her house	[3.3596…]	1
…			
2020	An **atom** is made up of protons, neutrons, and electrons.	[3.0123…]	2

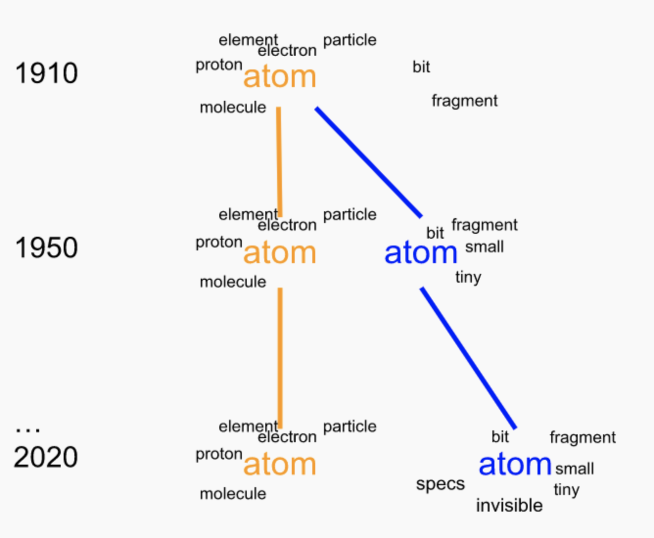
4. obtain word embeddings from time-specific mBERT at each time period. in visualisation, trace the different senses

period	sentence	time-specific word embeddings	sense
1910	There's not an **atom** of dirt in her house	[2.0968…]	1
1910	I couldn't find an **atom** of hatred in the sweet, innocent girl.	[2.0045…]	1
1910	An **atom** is made up of protons, neutrons, and electrons.	[4.1230…]	2
1910	Although containing an asymmetric carbon **atom** it has not been resolved.	[4.5312…]	2

Visualisation tools

final visualisation

*neighbours are made up



Visualisation tools

2 main ideas:

- Using the Python library Bokeh → Producing a simple cluster-like graph
- Using the Stardog API → representation of the RDF graph where an entity is the meaning of linked to its translation in several languages

Timeline

- end of November: understanding how to trace multiple senses in BERT; obtain corpora; further pre-train mBERT on two different periods of multilingual data; get program to generate quantified measurements of semantic change out of mBERT; prototype visualisation component.
- Mid-December: evaluation with existing benchmark and on historical events; finish implementing solution to the multiple senses problem
- End of December: finish training on all the time periods
- Rest of January: writing report

Thank you!

Question time