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 13 January 2023

Outline

1. Introduction

2. Work done so far
   - `scalable_semantic_shift` pipeline adaptation
   - visualization

3. Ongoing work and challenges
   - Measuring Semantic Shift
   - Evaluation Dataset
   - UI and Visualisation Issues

4. Conclusion

Introduction

Recap: Project Summary

- goal: tracing and visualizing diachronic semantic change using contextualized embeddings, by further pre-training m-BERT on an array of multilingual time-segmented corpora
- main update: adapting scalable_semantic_shift (`https://github.com/matejMartinc/scalable_semantic_shift`) - using cluster comparison to determine lexical semantic change numerically

Work done so far

scalable_semantic_shift pipeline adaptation

- Done so far: Corpus choice, preprocessing, model training, embeddings extraction...
- This repo provides a lot of tools useful for our project, as it was a very similar project in scope and goal, but with some key differences. (Mono-model rather than separate models per time slice; research focus with hardcoded defaults, rather than a toolkit/app focus with UI, ...)
- As we are now using SemEval, we plan to have a 2nd look at `get_embeddings_scalable_semeval.py`.

Relevant set of scripts from this repo that we are wrapping and adapting:

```
build_coha_corpus.py -- done
fine-tune_BERT.py -- done
get_embeddings_scalable_semeval.py -- SKIP
get_embeddings_scalable.py -- done
measure_semantic_shift.py -- partially done
evaluate.py
interpretation.py
```

Visualization

- **Tool:** Bokeh for Python, easy to set up and use
- Read the outputs from *measure_semantic_shift.py* and feed them to the visualization

`http://127.0.0.1:5000/analogy/awful/results`

Ongoing work and challenges

Measuring Semantic Shift

Overview:

- Baseline: Wasserstein Distance and Jensen-Shannon Divergence
- Input: Word embeddings and sentence references
- From: Pickle files generated by `scalable_semantic_shift`

Works to be done:

- Train the model to generate sufficiently large embeddings
- Merge the slice-based embeddings
- Examine the word sense loss/gain over the slices
- Rearrange the embeddings with Proscrutes Regression

Evaluation Dataset

For comparable and automatic evaluation, we will use the SemEval 2020 task 1 datasets (english and german), which come with the respective training data and a manually annotated test set.

- **training**: preprocessed corpora of time periods C_1 and C_2
  - English: COHA
  - German: news data - DTA for C_1 (1800-1899) and Berliner Zeitung & Neues Deutschland for C_2 (1946-1990)
- **test**: a list of target words with a number which indicates the amount of semantic change, that allows for comparison of amount of change across words

Web App Issues

How to manage user's requests ?

- Size available on Grid5k: $\sim$ 25 Gb
- Size of each pytorch model: $\sim$ 680 Mb
- Generating all the pickle files for all time periods for all words would be too heavy

$\rightarrow$ Generate the pickle files for a determined list of words and make this list available to the user.

Workflow

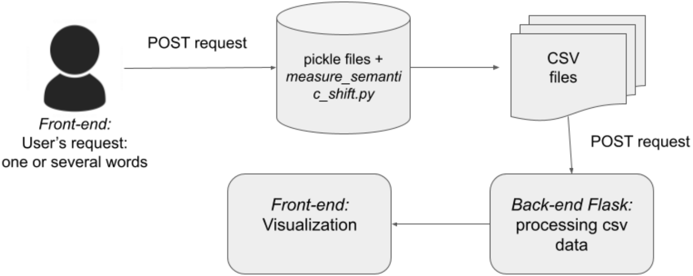

Figure: User's request workflow

Conclusion

Conclusions and next steps timeline

- Completed: *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.*
- TBD This week (before Thu 19 Jan): re-train models on the larger corpus slices (from SemEval); evaluation with existing benchmarks (SemEval); implement alignment if necessary; connect real results data into the visualisation UI; decide on practical trade-off solutions for visualisation UI and implement them.
- TBD Next week (before Thu 26 Jan): tweak training to improve eval results; writing report
- TBD Fri 27 Jan: turn in report

Thank you!

Question time