Diachronic study of semantic analogies
Software project, group 5

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

Université de Lorraine

Monday 7th November, 2022

Outline

Introduction

- **Project goal:** A project related to the 5th proposed project on cross-lingual analogy. Instead of across related languages, we are interested in a diachronic study and determining the level of similarity of the same language (e.g. English) across different historical periods according to semantic analogies.

What we did so far

- Browsing and choosing datasets
- Running 3 different baselines
- More literature review on temporal word embeddings and temporal word analogies

Datasets

Potential datasets

Our first choice for English:

- DUKweb: Diachronic UK web
  - Compile resources from Internet Archive of contemporary British English from 1996 to 2013 (granularity: years)
  - Contains co-occurrences matrices and word vectors for each year

  Other datasets from various languages:

- Datasets constructed from the SemEval Task:
  - RuSemShift (Rodiana and Kutuzov, 2021) for Russian
  - Kronos-it (Basile, Semeraro, Caputo, 2019) for Italian
  - DIACR-Ita (Basile et al., 2020) for Italian
  - LSCDiscovery (Zamora-Reina, Bravo-Marquez, Schlechtweg, 2022) for Spanish
  - ZhShfitEval (Chen, Chersoni, Huang, 2022) for Chinese Mandarin
  - NorDiaChange (Kutuzov et al., 2022) for Norwegian

Literature review

Temporal Word Analogies (TWA)

"A temporal word embedding is a pair of words which occupy a similar semantic space at different points in time." (Szy17)

Method used for discovering TWA :

1. **Train independent Vector Space Models**: 1 VSM = 1 period of time. Uses Word2Vec to obtain the models.

2. **Transform all those VSMs into a unique common space**: Align different VSMs from the different periods. 3 methods are available :
   - **Non-random initialization**
   - **Local linear regression**
   - **Orthogonal procrustes**

3. **Compare a word's vector from different VSMs**: To define patterns of change over time.

Baseline Work

Szymanski et al. 2017 (TWAPY)

- 1987 : reagan :: 1997 : clinton (US presidency)
- 1987 : mitterrand :: 1997 : mr_chirac (French presidency)
- 1987 : gorbachev :: 1997 : mr_yeltsin (From Soviet to Russian presidency)
- 1987 : soviet :: 1997 : russian (Historical event based on entities)
- 1987 : gemayel :: 1997 : denktash (Lebanese presidency, Hrawi ~~denktash~~)
- 1987 : icbm :: 1997 : warheads (Strategic changes in modern warfare?)
- 1987 : philippines :: 1997 : united_states (Political crisis)
- 1987 : internet :: 1997 : ... (OOV)

Arefyev and Zhikov 2020 (BOS AggloSil)

- N Arefyev, V Zhikov. BOS at SemEval-2020 Task 1: Word Sense Induction via Lexical Substitution for Lexical Semantic Change Detection. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, 171-179. Bib: (AZ20) Links: web, pdf

- Their main repo:
  `https://github.com/DeadBread/BOS_AggloSil`

- Corresponding visualization tool (web app):
  `https://github.com/DeadBread/SCD_WSI_tool/`

- BOS+AggloSil+DC: "WSI [word sense induction] is performed by clustering bag-of substitutes (BOS) vectors with agglomerative clustering employing silhouette score to select the number of clusters (AggloSil). Finally, we search for a decision cluster (DC), i.e. a cluster that has large number of occurrences from one corpus and small from another, and predict semantic change if such cluster exists."

BOS AggloSil contd.

- Goal: Get the codebase working, before trying to decide to adapt it for our purposes or not
- We were able to familiarize ourselves with the codebase and run their example commands. Challenges included:
  - codebase is old, and requires specific versioning (resolved by testing different python versions, creating a requirements.txt and conda environment with each problem ironed out);
  - codebase is also not well documented (explored the code with various tools, including creating call graph diagrams with code2flow);
  - not having enough RAM to run the code (resolved with teamwork, Averie was able to run it).

Rosin et al., 2017

- uses analogy with *entity-relation* pairs to look for change in semantic *relatedness* rather than *similarity*

    - *'in which time period were the words Obama and president maximally related?'*
    - Obama-president vs. Obama-Trump
    - relies on an additional relational corpus (YAGO2)

Similarity between 'Obama' and 'President'

- issues:
    - available relations are rather limited, so we decided it won't be the main component of our project
    - the purpose is more about temporal querying (factual) rather than exploring the semantics of words
    - haven't been able to run it yet due to missing NYT dataset
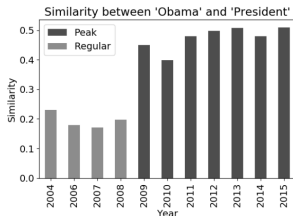
Conclusion

Conclusions and next steps

- Summary
  - we decide to use domain-general, longitudinal ($\sim$ 1 century+ wide) datasets for multiple languages
  - we would like to build on the codebase from Szymanski et al (2017) since it is the most similar for our purpose
  - we explored alternative ways to use analogies (semantic relations) and to obtain semantic representations (BOS AggloSil)
- Next steps
  - since a lot of the literature and the codebase are kind of old (-2017), we plan to consult the SemEval shared task (2020) to experiment with various technical aspects to improve the results (how to generate embeddings, how to compare vector representations across time periods..)
  - adapt the codebase to our datasets
- Remaining questions
  - understanding the (dis)advantages of using pretrained embeddings vs raw text datasets

[AZ20] Nikolay Arefyev and Vasily Zhikov, *BOS at SemEval-2020 task 1: Word sense induction via lexical substitution for lexical semantic change detection*, Proceedings of the Fourteenth Workshop on Semantic Evaluation (Barcelona (online)), International Committee for Computational Linguistics, December 2020, pp. 171–179.

[Szy17] Terrence Szymanski, *Temporal word analogies: Identifying lexical replacement with diachronic word embeddings*, Barzilay, R., Kan MY.(eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, 2017.

Thank you!

Question time