Diachronic study of semantic analogies

Software project, group 5

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

Université de Lorraine

Monday 24th October, 2022

Outline

1. Introduction

2. Literature review

3. Datasets

4. Demo

5. Conclusion

Introduction

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.
- For reference, *Proposed Project V – Cross-lingual analogy* was:

  > *Clustering of Semantic Analogies for Inter-Language Similarities: Recent works have allowed to retrieve language families from the transfer of morphological analogies between languages [9]. We wonder if we can find similar emerging patterns from multilingual or cross-lingual semantic analogies.*
  >
  > *Project goal: use clustering methods to explore, visualize and analyze the latent space of deep learning models of morphological analogy.*

Literature review

Previous works

What has been done:

- Diachronic Word Embeddings
- Cross-lingual Word Embeddings
- Diachronic and Synchronic Meanings

Previous works

Diachronic Word Embeddings

- Embeddings created from diachronic texts
- The texts should be associated to regular contexts

Cross-lingual Word Embeddings

- Embeddings created from multilingual texts
- Focus on a certain list of languages
- The contents of texts should be related to each other across languages

Diachronic and Synchronic Meanings of Words

- Changes of semantic meaning may:
  - Be synchronic or diachronic
  - Vary across fields and professions (Synchronic)
  - Vary across societies (Synchronic)
  - Refer to social/cultural events and differences (might be both?)
- → Posing challenges in choosing/constructing the datasets

Datasets

Factors to consider

- Language: how many languages do we want to consider
- Granularity: semantic change across 100 years or 1 month?
  - Kutuzov et al, 2018:
    smaller time spans → socio-cultural semantic changes
    longer time spans → linguistic semantic changes
- Domain: effects of using data from different domains

Potential datasets

- Google Books Ngrams (1800s-2000s)
    - PROS: multilingual (8 languages), wide time span
    - CONS: only up to 5-grams, not full sentences
- Twitter data (2006-present)
    - PROS: fine-grained granularity
    - CONS: more variation in word forms, difficult to evaluate?
- COHA (Corpus of Historical American English) (1820s-2010s)
    - PROS: test-set available, potential synchronic comparisons (eg. different variations of English, cross-domain)
    - CONS: monolingual

Evaluation datasets

- Test Data for SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection (EN, DE, SW, LA): manually annotated test set for semantic change over two time periods C1 and C2
  - for English C1 = 1810-1860, C2 = 1960-2010
  - binary semantic detection: did *'plane'* gain / lose a sense between C1 and C2?
  - graded semantic detection: how much did *'plane'* change between C1 and C2?
- other evaluation methods
  - identify time period given a sentence
  - given embeddings from C1 and C2, predict semantic change in C3
  - downstream tasks such as QA

Demo

Demo

https://example-app-analogies.herokuapp.com/index/
http://127.0.0.1:5000/

Conclusion

Conclusions and next steps

- What we've done so far
  - Overview of the research space so far, including a literature review of previous work done; potential datasets; potential codebases
  - Planning for what type of final tool we want to produce
- Challenges
  - Dataset: ensuring we have sufficient and quality data for the task. This can particularly be a problem for older variants of a language, and especially outside of the very high-resource languages.
  - Doubts about efficacy of vector offset for analogical reasoning: *"If the formulation of vector offset excludes the source vectors, it will appear to work for the small original dataset, where much of its success can be attributed to basic cosine similarity. But it will fail to generalize to a larger set of linguistic relations."* (Rog19)
- Next steps
  - Settling on (a) dataset(s)
  - Settling on (a) codebase(s) to use
  - After these decisions, we can start writing the code and running the experiments

References I

[Rog19] Anna Rogers, *On word analogies and negative results in nlp*, Jul 2019.

Thank you!

Question time