Detecting semantic shifts in biomedical literature through an intra-year and inter-year approach

This manuscript (permalink) was automatically generated from greenelab/word_lapse_manuscript@d7dec8b on March 30, 2022.

Authors

- **David Nicholson**
  0000-0003-0002-5761 · danich1 · dnicholson329
  Genomics and Computational Biology · Funded by Grant XXXXXXXX

- **Faisal Alquaddoomi**
  XXXX-XXXX-XXXX-XXXX · janeroe
  Center for Health Artificial Intellegence (CHAI)

- **Vincent Rubinetti**
  XXXX-XXXX-XXXX-XXXX · janeroe
  Center for Health Artificial Intellegence (CHAI)

- **Casey S. Greene**
  XXXX-XXXX-XXXX-XXXX · cgreene
  Department of Something, University of Whatever; Department of Whatever, University of Something

Abstract

Introduction

1. Biomedical language is constantly changing
    1. Scientists make new discoveries
    2. Old technologies are being revamped
    3. More reasons for language changes
2. Diachronic studies aims to capture these types of changes through time
    1. Insert papers for language changes in general
        1. doi:10.1007/s00799-019-00271-6
        2. doi:10.1142/9789811232701_0011
        3. arxiv:1605.09096
        4. arxiv:1806.03537
        5. arxiv:1606.02821
        6. doi:10.18653/v1/N18-1044
3. Gap in knowledge:
    1. work has focused majorly on non-biomedical text
    2. biomedical related work is only on abstract and titles
    3. Work has yet to focus on individual tokens and themes.
4. Goal is to examine longitudinal trends for biomedical term changes.

Don’t forget in here your methodological contribution to estimate uncertainty with variable-size training datasets in each year by training multiple models and examining variability across them.

Methods

1. Pubtator Central
    1. breif description about the dataset
    2. talk about how it contains entities tagged
2. Word2vec Model
    1. training parameters
    2. Use 10 models for each year
    3. min cutoff is 10
3. Orthogonal Procrustes - to align models onto year 2021
    1. Allows for the models to be directly compared
4. Determining semantic change
    1. Cosine metric to determine difference between words
    2. Scaf ratio method to model temporal changes
    3. CUSUM to actually detect change throughout the years
5. Umap visualization
    1. Explain why aligned umap - preserves local and global structure
    2. mention parameters for aligned umap

Results

One potential results panel: comparing results with abstract and full text (since you say above full text is not so often used).

1. Figure 1 - Visual to show that alignment works along intra year variation (Procrustes Validation)
2. Umap panel for an individual year to show models are separated
   I would start with the simplest possible way that you can show things, and only bring in more complexity like UMAP when necessary. I'm guessing that you would be able to show this with just PCA or similar, unless I'm not thinking correctly.
3. Umap panel to show the same year after models have been aligned
4. Umap panel to show across years (inter-year varation)
    1. might not have this as a figure given the aligned umap results, but we shall see
5. Figure 2 - CUSUM validation
   I think you might want another figure here that shows how you can use the variability in distance to address the challenges of variable training set size / model uncertainty. Let's see how things shake out as you're writing.
6. Table of found timepoint changes using this algorithm
7. Highlight pandemic as positive control
8. Nearest Neighbors upset plot ^^
9. Might look into lung cancer or other form of cancer results
10. Figure 3 - Website walkthrough for the work done here
11. Website screenshots
    The figures that make up the website might be nice to show/explain individually - not as screenshots but as actual figures (remember the PLOS situation where using a screenshot was an issue we had to work through).
12. Basically a walkthrough of the website and how a user can operate the web resource (similar to preprint similarity search)

Discussion and Conclusion

1. We provided a website resource to allow users to see toekn association changes
2. Word2vec is unstable and we implemented an approach to account for that variation
3. Constructs groundwork for future research into token changes
4. Will implement biorxiv and other preprint resources as a next step.

References