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@cef07c8 on May 18, 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

Biomedical Corpora Examined

Pubtator Central

Pubtator Central is an open-access resource containing annotated abstracts and full-text annotated with entity recognition systems for biomedical concepts [1]. The methods used are TaggerOne [2] to tag diseases, chemicals, and cell line entities, GNormPlus [3] to tag genes, SR4GN [4] to tag species, and tmVar [5] to tag genetic mutations. We initially downloaded this resource on December 07th, 2021, and processed over 30 million documents. This resource contains documents that date back to the pre-1800s to the year 2021; however, due to the low sample size in early years, we only used documents published from 2000 to 2021. The resource was subsequently updated with documents from 2021. We also downloaded a later version on March 09th, 2022, and merged both versions using each document's doc_id field to produce the corpus used in this analysis. We divided documents by publication year and then preprocessed each using spacy's en_core_web_sm model [6]. We replaced each tagged word or phrase with its corresponding entity type and entity id for every sentence that contained an annotation. Then, we used spacy to break sentences into individual tokens and normalized each token to its root form via lemmatization. After preprocessing, we used every sentence to train multiple natural language models designed to represent words based on their context.

Biomedical Preprints

BioRxiv [doi:10.1101/833400v1?] and MedRxiv [7] are repositories that contain preprints for the life science community. MedRxiv mainly focuses on preprints that mention patient research, while bioRxiv focuses on general biology. We downloaded a snapshot of both resources on March 4th, 2022, using their respective Amazon S3 bucket [8,9]. This snapshot contained 172,868 BioRxiv preprints and 37,517 MedRxiv preprints. These resources allow authors to post multiple versions of a single preprint. To prevent duplication bias, we filtered every preprint to its most recent version and sorted each preprint into its respective posted year. Unlike Pubtator Central, these filtered preprints do not contain any annotations. Therefore, we used TaggerOne [2] to tag every chemical and disease entity and GNormplus [3] to tag every gene and species entity for our preprint set. Once tagged, we used spacy to preprocess every preprint as described in our Pubtator Central section.

Constructing Word Embeddings for Semantic Change Detection

Word2vec [10] is a natural language processing model designed to model words based on their respective neighbors in the form of dense vectors. This suite of models comes in two forms, a skipgram model and a continuous bags of words (CBOW) model. The skipgram model generates these vectors by having a shallow neural network predict a word's neighbors given the word, while the CBOW model predicts the word given its neighbors. We used the CBOW model to construct word vectors for each year. Despite the power of these word2vec models, these models are known to differ both due to randomization within year and year-to-year variability across years [11,12,13,14]. To control for run-to-run variability, we examined both intra-year and inter-year relationships. Each year, we trained ten different CBOW models using the following parameters: vector size of 300, 10 epochs, minimum frequency cutoff of 5, and a window size of 16 for abstracts. Every model has its own unique vector space following training, making it difficult to compare two models without a correction step. We used orthogonal Procrustes [15] to align models. We aligned all trained CBOW models for the Pubtator Central dataset to the first model trained in 2021. Likewise, we aligned all CBOW models for the BioRxiv/MedRxiv dataset to the first model trained in 2021. We used UMAP [16] to visually examine the aligned models. We trained this model using the following parameters: cosine distance metric, random_state of 100, 25 for n_neighbors, a minimum distance of 0.99, and 50 n_epochs.

Detecting semantic changes across time

Once word2vec models are aligned, the next step is to detect semantic change.
Semantic change events are often detected through time series analysis [17]. We constructed a time series sequence for every token by calculating its distance within a given year (intra-year) and across each year (inter-year). We used the model pairs constructed from the same year to calculate an intra-year distance. Then, we calculated the cosine distance between each token and its corresponding counterpart for every generated pair. Cosine distance is a metric bounded between zero and two, where a score of zero means two vectors are the same, and a score of two means both vectors are different. For the inter-year distance, we used the Cartesian product of every model between two years and calculated the distance between tokens in the same way as the intra-year distance. Following both calculations, we combined both metrics by taking the ratio of the average inter-year distance over the average intra-year distance. Through this approach, tokens with high intra-year instability will be penalized and vice-verse for more stable tokens. Along with token distance calculations, it has been shown that including token frequency improves results compared to using distance alone [18]. We calculated token frequency as the ratio of token frequency in the more recent year over the frequency of the previous year. Then, we combined both the frequency and distance ratios to make the final metric.

Following time series construction, we performed change point detection, which is a process that uses statistical techniques to detect abnormalities within a given time series. We used the CUSUM algorithm [19] to detect these abnormalities. This algorithm uses a rolling sum of the differences

between two timepoints and checks whether the sum is greater than a threshold. A changepoint is considered to have occurred if the sum is greater than a threshold. We used the 99th percentile on every generated timepoint as the threshold. Then, we ran the CUSUM algorithm using a drift of 0 and default settings for all other parameters.

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

1. **PubTator central: automated concept annotation for biomedical full text articles**
Chih-Hsuan Wei, Alexis Allot, Robert Leaman, Zhiyong Lu
Nucleic Acids Research (2019-05-22) https://doi.org/ggzfsc
DOI: 10.1093/nar/gkz389 · PMID: 31114887 · PMCID: PMC6602571

2. **TaggerOne: joint named entity recognition and normalization with semi-Markov Models**
Robert Leaman, Zhiyong Lu
Bioinformatics (2016-06-09) https://doi.org/f855dg
DOI: 10.1093/bioinformatics/btw343 · PMID: 27283952 · PMCID: PMC5018376

3. **GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains**
Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu
BioMed Research International (2015) https://doi.org/gb85jb
DOI: 10.1155/2015/918710 · PMID: 26380306 · PMCID: PMC4561873

4. **SR4GN: A Species Recognition Software Tool for Gene Normalization**
Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu
PLoS ONE (2012-06-05) https://doi.org/gpq498
DOI: 10.1371/journal.pone.0038460 · PMID: 22679507 · PMCID: PMC3367953

5. **tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine**
Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, Zhiyong Lu
Bioinformatics (2017-09-01) https://doi.org/gbzsmc
DOI: 10.1093/bioinformatics/btx541 · PMID: 28968638 · PMCID: PMC5860583

6. **spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing**
Matthew Honnibal, Ines Montani
(2017)

7. **Medical preprint server debuts**
Jocelyn Kaiser
Science (2019-06-05) https://doi.org/gpxkkf
DOI: 10.1126/science.aay2933

8. **Machine access and text/data mining resources | bioRxiv** https://www.biorxiv.org/tdm

9. **Machine access and text/data mining resources | medRxiv** https://www.medrxiv.org/tdm

10. **Efficient Estimation of Word Representations in Vector Space**
Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean
arXiv (2013-09-10) https://arxiv.org/abs/1301.3781

11. **Factors Influencing the Surprising Instability of Word Embeddings**
Laura Wendlandt, Jonathan K Kummerfeld, Rada Mihalcea
arXiv (2020-06-05) https://arxiv.org/abs/1804.09692
DOI: 10.18653/v1/n18-1190

12. **Stability of Word Embeddings Using Word2Vec**
Mansi Chugh, Peter A Whigham, Grant Dick

AI 2018: Advances in Artificial Intelligence (2018) https://doi.org/gpxkkc
DOI: 10.1007/978-3-030-03991-2_73

13. **Evaluating the Stability of Embedding-based Word Similarities**
Maria Antoniak, David Mimno
Transactions of the Association for Computational Linguistics (2018-12) https://doi.org/gf39k8
DOI: 10.1162/tacl_a_00008

14. **Predicting Word Embeddings Variability**
Benedicte Pierrejean, Ludovic Tanguy
Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (2018)
https://doi.org/gh6qpc
DOI: 10.18653/v1/s18-2019

15. **A generalized solution of the orthogonal procrustes problem**
Peter H Schönemann
Psychometrika (1966-03) https://doi.org/dx77sz
DOI: 10.1007/bf02289451

16. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**
Leland McInnes, John Healy, James Melville
arXiv (2020-09-21) https://arxiv.org/abs/1802.03426

17. **Statistically Significant Detection of Linguistic Change**
Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, Steven Skiena
Proceedings of the 24th International Conference on World Wide Web (2015-05-18)
https://doi.org/ghcv6k
DOI: 10.1145/2736277.2741627

18. **Improving semantic change analysis by combining word embeddings and word frequencies**
Adrian Englhardt, Jens Willkomm, Martin Schäler, Klemens Böhm
International Journal on Digital Libraries (2019-05-20) https://doi.org/gpxkkd
DOI: 10.1007/s00799-019-00271-6

19. **Adaptive filtering and change detection**
Fredrik Gustafsson, Fredrik Gustafsson
Citeseer (2000)