Inference and Representation Fall 2017 Final Project Proposal

Zhi-Wei Li Alexander Rich Anselm Rothe

Topic models have been used successfully to capture structure in large text corpora and make these corpora more human-understandable (Blei, Ng & Jordan 2003). However, traditional topic models to not capture the temporal ordering of documents, and the way topics change over time. Dynamic topic models address this issue by allowing topic mixing proportions and the distribution of words in topics to change as a function of time (Blei & Lafferty 2006). In this project, we will apply dynamic topic modeling to 35 years of the Proceedings of the Cognitive Science Society to understand and visualize how the research topics of the society have changed over its history.

Data Set

The Proceedings of the Cognitive Science Society are a representative corpus of the ongoing research in the cognitive science community, and are available online for the 3rd through 38th annual conferences (1981-2016). While we have not yet counted the documents in the corpus, we expect it to be on the order of 10,000 documents, with each document about 6 pdf pages long. Older proceedings were scanned and processed using OCR, so there may be some issues with data quality for these earlier years. We will likely preprocess this data set using stemming and/or lemmatization techniques, and remove common words that do not carry interesting meaning.

Model

We will implement a dynamic topic model (DTM) as described by Blei and Lafferty (2006). As with traditional topic models, documents are assumed to be generated by a model in which topic proportions are first generated according to a Dirichlet distribution, words are assigned to topics according to a multinomial distribution based on the topic proportions, and then words are generated according to a multinomial distribution based on the assigned topic. In a dynamic topic model, topic proportions and word distributions within a topic are both allowed to drift over time according to a logistic normal distribution. The full graphical model is summarized in Figure 1. To fit the DTM to our data set, we will write the model as a directed graphical model using the Stan modeling language (Stan Development Team, 2017). We will then fit the model using Stan's built-in variational inference algorithm.

Evaluation

Following Blei and Lafferty, we will evaluate the model by fitting the model to all data up to a given year, and then testing its ability to predict the articles from that year. For comparison, we will also test traditional topic models trained on all previous years and on only the most recent previous year.

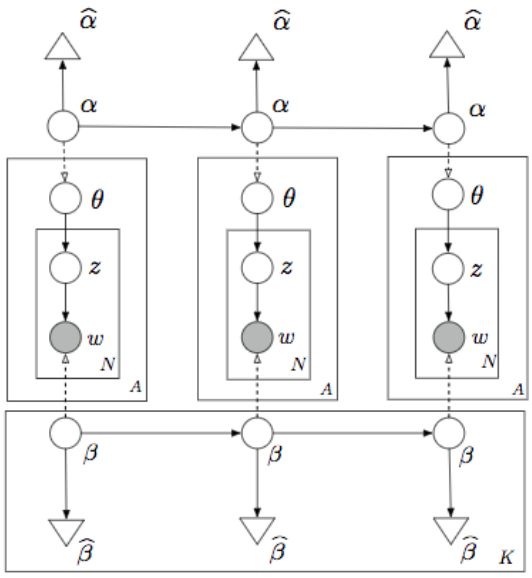

Figure 1: Full graphical model of a dynamic topic model (DTM).

Deliverables

Our goal is to use our model to understand how the field of cognitive science has changed over time, as attempted with traditional topic models by Cohen Priva & Austerweil (2015) for the journal Cognition (which focuses more squarely on cognitive psychology). To achieve this, we plan to visualize the models' topics and their change over time in an understandable manner. Building off the work of Blei and Lafferty (see Figure 2), we will visualize each topic by showing its shifting popularity (i.e., proportion) over time, top words over time, and a selection of article titles typical of the topic over time. We will also perform analyses to reveal which topics have changed the most and least in popularity over time, and which have changed most and least in the distribution of words within the topics.

Acknowledgements

We were granted an exception to form a team of 3 (instead of 2) students by Joan after the lecture on Oct 16.

References

Blei, D., & Lafferty, J.D. (2006). Dynamic topic models. Proceedings of the 23rd international Conference on Machine Learning. ICML '06, 113-120.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research. 993-1022.

Cohen Priva, U., & Austerweil, J.L. (2015). Analyzing the history of Cognition using Topic Models.

Stan Development Team. 2017. The Stan Core Library, Version 2.16.0. http://mc-stan.org

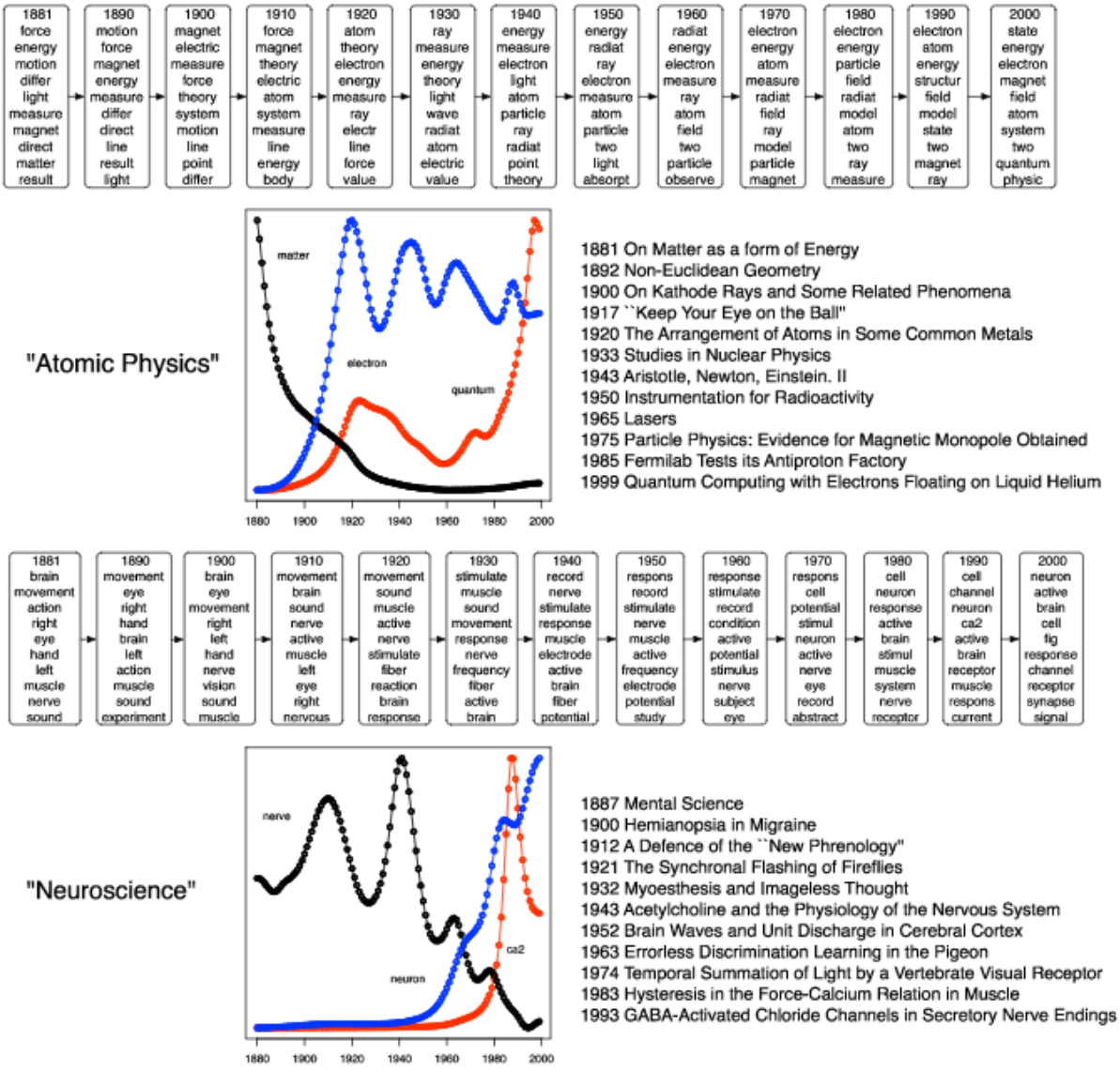

Figure 2: Trends uncovered by the DTM in Blei and Lafferty (2006).