Project Idea Relations: A Literature Review

Qitian Hu, Shicheng Liu, Yijie Yao

February 14, 2021

1 Minimal Requirement

A list of the main empirical results presented in the paper to be replicated [1] follows:

1. Figure 1 on **News Articles**
2. Figure 3 on **News Articles** (topic: Terrorism and Immigration) and **Research Papers** (topic: ACL)
3. Section 4.1, 4.2, Figure 5, Figure 6, Figure 7 and Table 1 on **News Articles** (topic: Terrorism and Immigration)
4. Section 4.3, Figure 8 and Figure 9 on **Research Papers** (topic: ACL)

Among these, the **Research Papers** are provided and can be downloaded. The **News Articles** needs further investigation (it seems on the project list guideline, replicating results on **Research Papers** would suffice for the minimum requirement. However, we believe it would be worthwhile replicating/exploring the **News Articles** as well.)

We were able to download the open-sourced code and have begun the first phase of replicating the results of the paper. In particular, we focused on replicating part of (4) of the above list. It seems simply re-running the provided 'example.sh' (no matter with 'num_ideas' equal to 50 or 100) file with the downloaded ACL dataset does not give the same result as the relationship graph in Figure 8 (seems like the output contains very different words as the ones mentioned in the paper - 'machine translation', 'sentiment analysis', 'word alignment', 'discourse (coherence)', and 'rule, forest methods'). Thus, this will need further inspection. We'll continue the investigation of potentials reasons for these different results. We will reach out to Professor Tan, and if this issue remains unsolved.

2 Related Ideas and Methodological Extension

There are many possible routes we could take to further explore the topics of idea relations, and in this section we two directions we would like to explore: one is through different approaches to identifying ideas; the other is through implementation on different corpora.

2.1 Different Approaches for Identifying Ideas

The original paper provides us a foundational framework for analyzing the relation between ideas. The proposed framework used the standard topic modeling (LDA) to identify ideas. We identified several points of weakness and brainstormed potential methods for improvement:

1. There is a degree of randomness involved in this process and the topics we identify rely on the training process and random seed.

2. We cannot incorporate prior knowledge on the corpus and the ideas in it. Since topic modeling is an unsupervised method, we cannot guarantee to have some topics that we're interested in.

3. For those corpora that their structure, focus, and topic change a lot through time, the current mechanism might be unable to capture this shift. For example, for certain corpus, we may have some topics that are only present in the early times and some present only in the late times. The existing framework might not able to highlight such change.

Here is a running list of alternative approaches and potential research questions associated with them:

1. We can conduct topic modeling with neural networks. This is one of Chenhao's later works that is also implemented on the same corpus of article regarding US immigration. This paper provides us with an instruction to incorporate metadata (like date, author, and etc) into topic modeling. [2]
   Possible Research Questions: Do the stated advantages of incorporating metadata in idea identification show significance when implementing with idea relation? How do we measure and show such advantage compared to the original paper that did not employ this method?

2. We can consider including the framework of interactive topic modeling. [3] This framework will allow us to encode their feedback easily and iteratively into the topic models. This method will also be able to add contextual information to the documents and perform more directed topic modeling. There are several seminal papers on this method and we will explore them to see which one is the most ready-to-use for our purpose. [7] [6]
   Possible Research Questions: How are the difference in resulting topics identified through interactive topic modeling and original method? Again, how do we measure such improvement and implement on different copora.

2.2 Implementation on Alternative Corpora

2.2.1 US Economic News Articles

We are interested in how idea relations evolve in the news that reflect the US economy. This data set provides us news article that is regarded relevant to the US economy, spanning from 1951 to 2014. [4] The dataset was originally used for sentiment analysis. By extending the case studies of the original paper to this alternative dataset, we are interested see if certain degrees of cycles can be captured with idea relations which correspond to the actual macroeconomics. We will also be able to confirm some economics intuitions of how certain ideas cooperate and compete based on the included financial news. We are curious if the relationship of the tryst, friendship, head-to-head, and arms-race will still hold.

Possible Research Questions: The article for the data set is much more concise than that of the original paper. We want to investigate how the length of article will impact our implementation. We are also interested in learning if the four prominent idea relationship identified in the original paper still carry significance. Do we need to define new relationships between ideas for cycles for example?

2.2.2 People's daily

We find that the corpus People's Daily is also open and available online. People's Daily is a China's national newspaper run exclusively by the Chinese government. Established in 1946, it is one of the most accurate representations of the government's self-perception, policy, and ideology. It would be worthwhile to use the framework of idea relations to see the change of popular ideas and arguments in the history of modern China. The data set is composed of the People's daily from 1950 to 2010, across 60 years. We will be able to obtain the preprocessed dataset of People's daily based on Li and Hovy's paper. [5]

Possible Research Questions: One thing that might be worth noting is that while the existing applications of this framework are mostly about a collection of texts produced by different entities, People's Daily is written by one centralized institute. Should we incorporate other entities and newspapers? Can we still use the regular interpretations of two ideas "competing" with each other as in Chenhao's original paper? Also through the implementation of this data set, we will become more familiar with techniques of the process of the Chinese language, and we will be asking questions about how the framework proposed by the original paper differs when we change from the English language to Chinese.

References

[1] Chenhao Tan, Dallas Card, Noah A. Smith. "Friendships, Rivalries, and Trysts: Characterizing Relations between Ideas in Texts"

[2] Card, Dallas, Chenhao Tan, and Noah A. Smith. "Neural models for documents with metadata." arXiv preprint arXiv:1705.09296 (2017).

[3] Demszky, Dorottya, et al. "Analyzing polarization in social media: Method and application to tweets on 21 mass shootings." arXiv preprint arXiv:1904.01596 (2019).

[4] Economic News Article Tone and Relevance https://data.world/crowdflower/economic-news-article-tone

[5] Jiwei Li and Eduard Hovy "Sentiment Analysis on the People's Daily", https://www.aclweb.org/anthology/D14-1053.pdf

[6] https://arxiv.org/pdf/1206.3298.pdf

[7] https://dl.acm.org/doi/10.1145/1143844.1143859