NLP Final Project: 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** is 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 minimum requirement. However, we believe it would be worthwhile replicating/exploring the **News Articles** ones as well.)

We were able to download the open-sourced code and have begun replicating 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 donwloaded 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 nees further inspection. We'll continue exploration and if this remains unsolved, we'll reach out to Chenhao.

2 Related Ideas and Methodological Extension

There are a lot of ways by which we could build on the existing paper, and in this section I outline several directions we would like to explore.

2.1 Other ways of finding ideas

While the original paper provides us an interesting framework of analyzing the relation between ideas, it has used the standard topic modeling (LDA) to identify ideas. We think there are several weak points that it could be modified:

1. There are 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 be sure that we will guaranteed to have some topics that we're interested in.
3. For those corpus that their own structure, focus, and topic change a lot through time, it might be unable to capture this shift, and we may have some topics that are only present in the early times and some present only in the late times.

With the help of Chenhao, we identified several possible alternatives that we will explore.

1. Topic modeling with neural networks. This is Chenhao's own work and allows us to incorporate metadata (like date) into topic modeling. [2]
2. Interactive topic modeling. [3] This method seems to be able to add contextual information to the documents and perform more directed topic modeling. There are a number of seminal papers on this method and we will explore them to see which one is the most ready-to-use for our purpose. [5] [4]

2.2 Other Corpus

In addition to the corpus in the original paper, we find that the corpus People's Daily is also open and available online. People's Daily is a national newspaper in China, and is directly controlled by the Communist Party. Established in 1946, it is probably the most authoritative representation of the government's self-perception, policy, and ideology. We think it would be interesting to use the framework of idea relations to see the change of popular ideas and arguments in the history of modern China.

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?

3 Possible Research Questions

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

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