Characterization of Idea Relations in Text: Investigation with Topic Modelling and Structural Topic Modelling

Shicheng Liu, Yijie Yao, Qitian Hu
Team: Mission Inspiring God
The University of Chicago
{shicheng2000,yjyao, jasonhu}@uchicago.edu

Abstract

We present a study on characterizing relations between ideas in text, inspired by Tan et al. (2017). We successfully replicate results on the ACL dataset of Tan et al. (2017). We then propose a new method to generate topics - structural topic modelling - and discuss its similarities and differences with raw topic modelling. We apply this framework to two new datasets - American Economic Review and People's Daily. We discuss and analyze the resulting trends and hypothesis.

1 Introduction

1.1 Overview

"*Ideas exist in the mind, but are made manifest in language, where they compete with each other for the scarce resource of human attention*" (Tan et al., 2017). Indeed, it is natural to ask the following questions:

1. What are the "ideas" presented in text and how do we capture them?
2. How do we think about their relationships with each other?

We attempt to seek out the answers utilizing tools from natural language processing. In this study, we consider representing ideas as topics obtained in an unsupervised fashion, using either topic modelling or structural topic modelling. Each document can then consist of one or more topics as determined (answering Question 1). We then apply a computational framework that attempts to capture how ideas quantitatively relate to each other in text (answering Question 2).

1.2 Contribution

The main contribution of our study, in addition to confirming the ACL-related results presented by Tan et al. (2017) in §3, is two-fold:

1. We propose a new method to generate ideas from text and we compare its outcome with the raw topic modelling method in §2.2 and §5;
2. We apply this modified computational framework, which utilize both raw topic modelling and structural topic modelling, on two new in-depth case studies. We identify and present new discoveries and hypothesis based on quantitative output in §4.

2 Computational Framework

In this section, we summarize the computational tools used to (1) identify ideas in text and (2) capture their relations.

2.1 Topic Modelling

Topic modeling is an unsupervised machine learning technique capable of scanning a set of documents and automatically clustering word groups and similar expressions that best characterize a set of documents. In this study, we utilize the LDA model given by (Blei et al., 2003). We will refer to this as the "raw topic modelling" in the following discussion.

2.2 Structural Topic Modelling

The raw topic modeling (raw TM) only adds topic and document as internal organization of the corpus, but in practice, we usually have a much larger range of metadata. Structural topic modeling or sTM (Roberts et al., 2013b) allows us to add document-level metadata on which topical prevalence or topical content is of interest. The main idea is to use generalized linear models as priors and then condition on metadata of the document.

Based on three variants of LDA (Blei et al., 2003), correlated topic model or CTM (Blei et al., 2007), the Dirichlet-Multinomial Regression topic

model (Mimno and McCallum, 2008)) and the Sparse Additive Generative (Taddy, 2013) topic model.

The logistic normal prior for topics in the standard CTM is replaced by a logistic-normal linear model. The design matrix for the covariates X allows for flexible forms of the covariates that the prior could be conditioned on.

The distribution over words is replaced with a multinomial logistic function such that a token's distribution is jointly determined by topic, covariates, and topic-covariate interaction. We adopt the R implementation provided by the authors (Roberts et al., 2019).

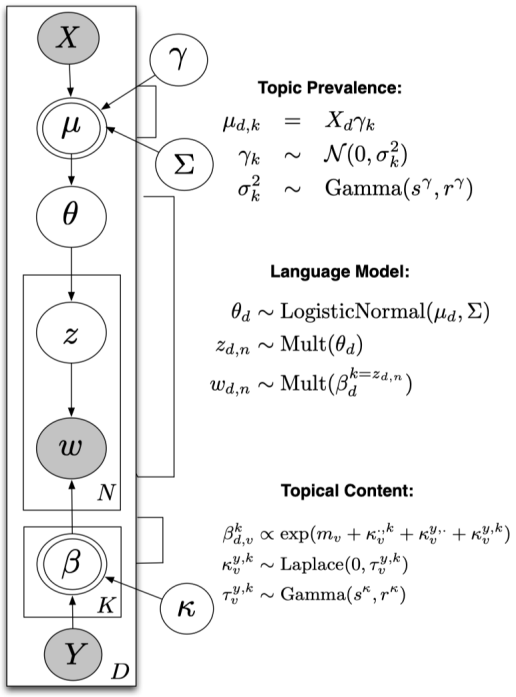

Figure 1: Plate Diagram for the Structural Topic Model

The generative process of sTM could be specified in three steps. First, draw the document-level attention to each topic from a logistic-normal generalized linear model based on a vector of document covariates X_d.

$$\vec{\theta}_d \mid X_d\gamma, \Sigma \sim LogisticNormal\left(\mu = X_d\gamma, \Sigma\right)$$

where X_d is a 1-by- p vector, γ is a p -by- $K-1$ matrix of coefficients and Σ is $K-1$ -by$K-1$ covariance matrix. Second, given a document-level content covariate y_d, form the document-specific distribution over words representing each topic (k) using the baseline word distribution (m), the topic specific deviation $\kappa_k^{(\mathrm{t})}$, the covariate group deviation $\kappa_{y_d}^{(\mathrm{c})}$ and the interaction between the two $\kappa_{y_d,k}^{(\mathrm{i})}\Big)$

$$\beta_{d,k} \propto \exp\left(m + \kappa_k^{(\mathrm{t})} + \kappa_{y_d}^{(\mathrm{c})} + \kappa_{y_d,k}^{(\mathrm{i})}\right)$$

m, and each $\kappa_k^{(\mathrm{t})}$, $\kappa_{y_d}^{(\mathrm{c})}$ and $\kappa_{y_d,k}^{(\mathrm{i})}$ are V -length vectors containing one entry per word in the vocabulary. When no convent covariate is present β can be formed as $\beta_{d,k} \propto \exp\left(m + \kappa_k^{(\mathrm{t})}\right)$ or simply point estimated (this latter behavior is the default).

Third, for each word in the document, $(n \in 1, \ldots, N_d)$ Draw word's topic assignment based on the document-specific distribution over topics.

$$z_{d,n} \mid \vec{\theta}_d \sim Multinomial\left(\vec{\theta}_d\right)$$

Conditional on the topic chosen, draw an observed word from that topic.

$$w_{d,n} \mid z_{d,n}, \beta_{d,k=z_{d,n}} \sim Multinomial\left(\beta_{d,k=z_{d,n}}\right)$$

2.3 Characterization of Idea Relations

Given a set of corpora, the above two methods will identify each corpus with a list of ideas. We then need a methodology to rigorously examine the relations between these ideas. One main contribution of Tan et al. (2017) is the proposal of a novel method to accomplish exactly this. The authors focus on two dimensions to characterize each relation between two ideas:

1. cooccurrence reveals to what extent two ideas tend to occur in the same contexts;

2. similarity between the relative prevalence of ideas over time reveals how two ideas relate in terms of popularity or coverage.

where the first dimension is captured by empirical pointwise mutual information (PMI) and the second by correlation between normalized document frequency of ideas. For details on this computation, please refer to the original paper.

In the presence of these two dimensions, each relation between two ideas can then be put into one of the four categories below:

1. **Friendship**: correlated over time, likely to cooccur

2. **Head-to-head**: anti-correlated over time, unlikely to cooccur

3. Tryst: anti-correlated over time, likely to cooccur

4. Arms-race: correlated over time, unlikely to cooccur

We employ the same framework to capture relations between ideas in this study.

3 Result Replication

In Tan et al. (2017), the authors explored the ACL dataset consisting of papers from ACL, NAACL, EMNLP, and TACL from 1980 to 2014. This dataset consists of 4.8K papers, of which a processed version is made available.

Using the open-sourced code and dataset in the initial paper, we are able to replicate results that are very similar to those presented in the paper.

Ori.	Rep.	First	Second
		Friendship	
1	1	word alignment	machine translation
		Arms-race	
1	1	sentiment analysis	machine translation
2	3	sentiment analysis	word alignment
23	38	rule,forest methods	discourse (coherence)
		Head-to-head	
1	3	discourse (coherence)	machine translation
7	Missing	discourse (coherence)	word alignment
38	Missing	rule,forest methods	sentiment analysis
		Tryst	
5	3	rule,forest methods	machine translation

Table 1: Result replication of the ACL-related data presented in Figure 8 of Tan et al. (2017). Column '**Ori.**' denotes the ranking of the relation as presented in Tan et al. (2017) while column '**Rep.**' denotes the ranking of the same relation in our experiment.

Running the **topic** method as described in Chenhao's work (Tan et al., 2017), we obtain a set of topics that are *associated* with particular topic words as determined by LDA (Blei et al., 2003). The code outputs the top 50 relations for each of the proposed idea relations: **Friendship**, **Head-to-head**, **Tryst**, **Arms-race**. Each of the two topics in these relations are represented by a set of words. We then need to manually name these topics and identify their associations.

Results of our experiment are shown in Table 1, which compares our result with the presented result. For details on which topic words are identified with what relations, see Table 2. Table 1 shows that for the most part, our results line up well with the ones presented in the original paper with slight differences (most notably on the two **Head-to-head** relations that were not found in the Top 50 results in our run). This could be due to randomness introduced in the topic modelling process but needs further investigation.

Notably, the association between identified topic and associated topic words also differs slightly with that presented in the original paper. In caption of Figure 8, Tan et al. (2017) identifies the **rule, forest methods** topic with 'rule, grammar, derivation, span, algorithm, forest, parsing, figure, set, string', which differs from the result of our experiment - 'rules, rule, rules., derivation, rules, ,set, figure, derivations, forest, synchronous'.

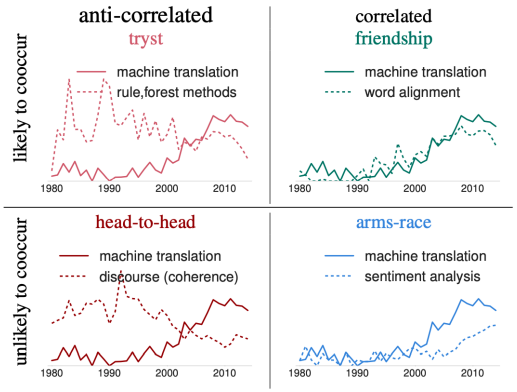

Figure 2: Highlights of certain relations as presented in Tan et al. (2017)

Fig. 3 shows our attempt at replicating Fig. 2 presented in the original paper. As the two figures show, the results, for the most part, match up very well. We believe the subtle differences are due to the inherent randomness introduced in the topic modelling process.

In conclusion, we have successfully replicated the ACL results presented in Tan et al. (2017). The presented results mostly match up with our experiment.

Identified Topic	Associated Topic Words
machine translation	translation, phrase, source, statistical machine, translation., system, mt, target, reordering, sentence
sentiment analysis	sentiment, opinion, positive, negative, polarity, reviews, review, words, aspect, product
word alignment	alignment, word, alignments, aligned, sentence, pairs, paraphrases, words, paraphrase, pair
discourse (coherence)	discourse, text, structure, relations, two, coherence, relation, focus, discourse., cue
rule,forest methods	rules, rule, rules., derivation, rules, , set, figure, derivations, forest, synchronous

Table 2: The set of topic words for each relation that appears in Table 1.

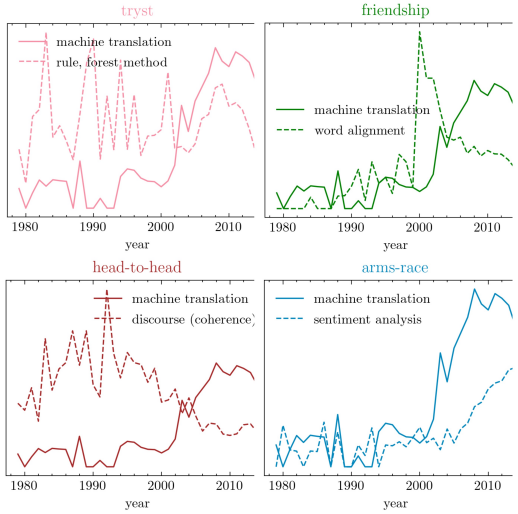

Figure 3: Replicated result from our experiment corresponding to Fig. 2

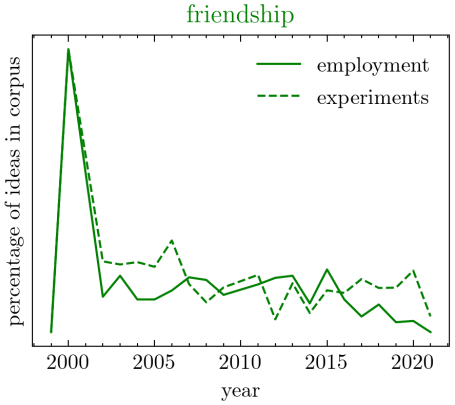

Figure 4: **Friendship** relation between 'employment' and 'experiments' in AER

4 Exploratory Studies

4.1 American Economics Review

American Economic Review (AER) is one of the most prestigious journals of Economics. AER journal is a representative of the intellectual history of the field. We are able to download over 4500+ journal articles from the 1999 to the 2021 from the AER official website. Here are some notable findings:

In Fig. 4, we see a **friendship** relation between "employment" and "experiments". The "employment" points to the topics including "employees, punishment, savings, contribution, default, monitoring, money, treatment, future, behavior"; the "experiments" points to the topics including "subjects, experiment, subject, rule, sub, experiments, rules, experimental, games, complexity". This friendship relation corresponds exactly to the relationship between the subject of economic research and its popular research methods. When the economists are studying the topics related to behaviors of employees under various motives within firms, it is usually conducted with some methods within the realm of experimental economics. So the **friendship** seems to be a fitting narrative here.

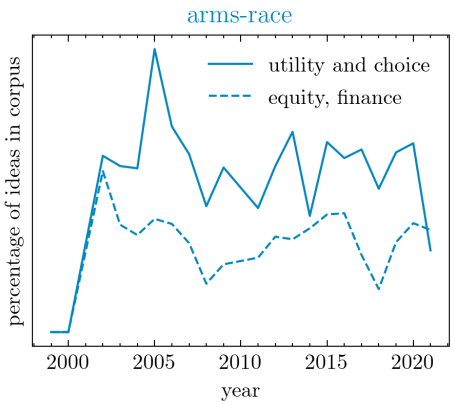

Figure 5: **Arms-race** relation between 'utility and choice" and 'equity, finance' in AER

In Fig. 5, we observe an **arms-race** between utility and choice versus equity, finance. The "utility and choice" here stands for consumer preference, utility, and decision theory. The "equity, finance" stands for the topic of financial economics, which includes the study of "equity, returns, asset, stock, return, investors, investment, wealth, mortgage". By observing these two topics to unlikely to cooccur, we can propose the hypothesis is that when a paper uses an approach of a theoretical model of consumer preference and utility, it is unlikely to come across with the topics of financial economics because the latter one increasingly more relies on empirical methods instead of theoretical reasoning.

Another interesting question we can ask is based on a **head-to-head** relation between "price auction bidding" and "ethnicity, immigrants" shown in 11 . The first one refers to the topic of "price, auction, and bidding mechanisms", and the second one refers to the topic of "ethnic groups and mobility of immigrants". We do not know the exact reason why these two topics are unlikely to cooccur and anti-correlated. One possibility is that the bidding and auction mechanism designed for the immigration path is never an object of study because such policies are losing favors and they are unlikely to happen together. But the question of why they are anti-correlated remained a myth.

We do want to point out although we can extract many notable relations from our results, the results could have been better. We saw left-end towards the years 1999 and 2000 seem to contain one or many outliers in many graphs. Because we have to obtain the AER dataset, we had to initially extract DOI number from the AER websites and download all the corresponding papers from JSTOR and other related sites. we had encountered some issues when converting PDFs to textual corpora with PDFMiner. Among 4500+ original papers, we eliminated about 1000 papers with surprisingly high percentage gibberish words. To further improve the results of the exploratory studies on AER, we would need to fine-tune our corpus preprocessing and reexamine our method of converting PDFs to txt files and keep greater number of journals with a high percentage of meaningful words.

4.2 People's Daily (REN MIN RI BAO)

People's Daily is the largest newspaper group in China, and is the official newspaper of the Central Committee of the Chinese Communist Party. It was founded in 1948 and represents the official ideology and policy of the Party. The corpus we we has about 1.3 million digital articles, ranging from 1948 until 2003.

4.2.1 Confirmation of Historical Knowledge

We discover that most of the fundamental shifts of the country's socio-intellectual history are accurately captured in the idea relations analysis.

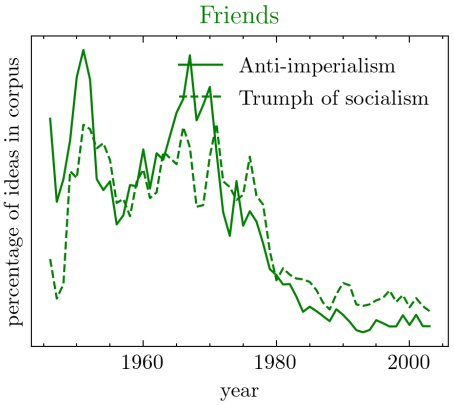

Figure 6: **Friendship** between "Anti-imperialism" and "triumph of socialism"

Friendship (Fig. 6): While China gradually transforms from an ideological nation to a more practical country aiming to for domestic economy, the prevalence of both the topic of anti-imperialism and the triumph of socialism declines. The **friendship** relation also demonstrates that the triumph of socialism rhetoric is closely connected to the anti-imperialism argument, and the superiority of one's own ideology is shown by belittle the opponent ideology.

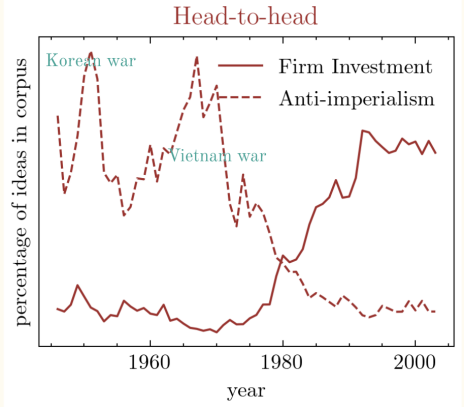

Figure 7: **Head-to-head** relation between "Anti-imperialism" and "firm investment"

Head-to-head (Fig. 7): Transformation of the country is a double process: the decline of the rev-

olutionary ideology and the rise of the economic rhetoric, and this figure perfects demonstrates this trend. Note that the two peaks of anti-imperialism topic correspond two important wars against the US, the Korean War and the Vietnam War; the official newspaper was also conducting an ideological campaign against the US.

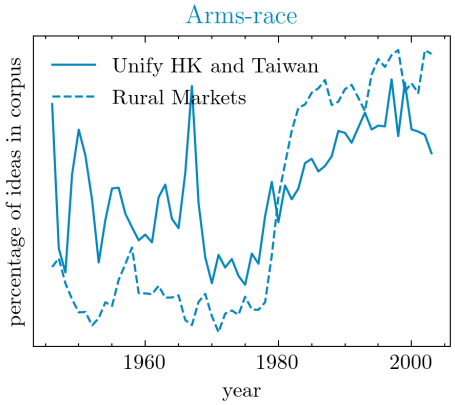

Figure 8: **Arms-race** relation between "Unify HK and Taiwan" and "rural markets"

Arms-race (Fig. 8): The topics of "Unify HK and Taiwan" and "rural markets" display similar trends in history but seem to be unrelated in terms of the driving force behind. This is accuratly captured by the lack of cooccurance. Unifying HK and Taiwan is an important theme throughout China's history, and it became important later in the timeframe because of specific events like Unification of HK and Taiwan Strait crisis and thus display a similar trend with the rural market topic, which is closely connected to the economic transformation of China.

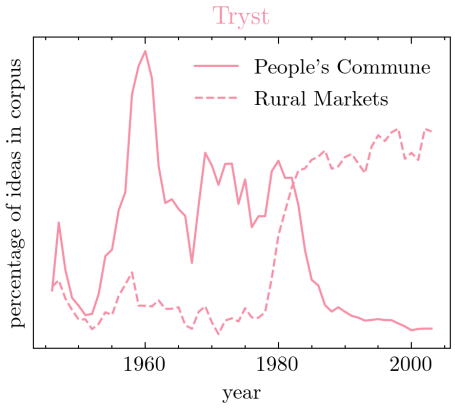

Figure 9: **Tryst** relation between "People's commune" and "rural markets"

Tryst (Fig. 9): "People's commune" and "rural markets" are two distinct narratives on the same issue of rural economic development, and this explains their tendency to cooccur. However, the former is a derivative of Communism while the latter the result of the market-oriented reforms, and as the country transforms, the time trends of the two topics are opposite.

4.2.2 New Hypothesis

The topic "diplomatic group visits" is consistently captured by the different algorithms and settings, and it is observed to be in a consistent **arms-race** with topics like "People's commune", "socialist construction", and other similar old, ideology-oriented ideas. It is unclear for now whether this is just driven by specific historical events or could be connected to more important themes in history, but hypothesis could be made for facilitate further exploration:

Hypothesis 1: diplomatic group in the Chinese context is an old-fashioned way of diplomacy and is sufficiently unrelated to domestic development.

Hypothesis 2: diplomatic group visits are usually reported in a special type of articles, and that type of articles are less published on People's Daily.

5 Comparison between sTM and raw TM

We choose to use People's Daily to compare the two topic models, because

1. The textual quality of the corpus is better
2. It is large and thick enough to demonstrate the difference
3. It has a long time frame so structural topic model might be effective

Although we use high-performance 128G-memory server, structural topic model fail to complete. We suspect it is due to the limit of the R language. Thus, we remove every word that appears in less than 10 documents or appear in more than half of all the documents. A brief comparison with the topics using the original corpus using raw topic model reveals that the quality of topics is not severely affected, and major historical themes are correctly identified.

We compare raw topic model and structural topic from 3 aspects. Firstly, we interpret the topics and check if important historical themes are captured. We discover that both models accurately capture the

themes of anti-imperialism, Chairman Mao and cultural revolution, socialism and communism, United Nations and diplomacy, Korea and Vietnam wars, market economy, firm and investment, etc. sTM tends to capture more detailed topics, like Israel and Palestine, India and Pakistan, and municipal constructions, while these topics were not captured by raw TM.

Secondly, to quantify the topic difference, we calculate the average number of characters of each word selected by the two topic models. Note that each Chinese word is composed by a number of characters, and usually the longer the word is, the more detailed and specific it is. The average word length for raw TM is 2.09 (std = 0.789), while that of sTM is 3.28 (std = 0.591). sTM tends to capture more detailed, specific words.

Third, we visualize the time range of documents in each topic. As the figure shows, while raw TM tends to capture topics that capture documents around 1976, sTM topics could vary in the mean year of related documents. Time range of Raw TM topics have standard deviation centered at 17 year, while sTM topics could be very specific to a period (low standard deviation) and vice versa.

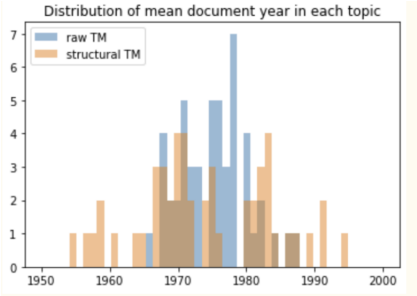

Figure 10: Distribution of mean document year for each topic

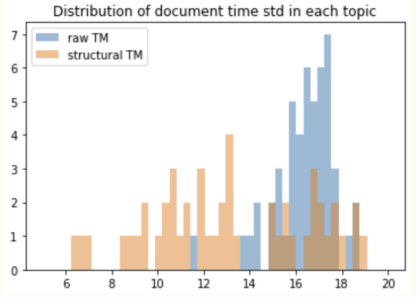

Figure 11: Distribution of document year standard deviation for each topic

To sum up, raw TM topics are more general and usually capture documents in a broader period. With metadata on the temporal change of prevalence of topics, sTM captures more specific topics of specific historical periods, but also long periods of time.

6 Related Work

We now present three strands of related studies in addition to what we have discussed.

6.1 Trends in ideas

Most previous studies before Chenhao's work focus on the trends of single ideas over an extended period of time. For example, Rule et al. (2015) presents a number of methods that reveal change in contents that masks continuity in the articulation of major governance task through analyzing State of Union Addresses from 1790 to 2014. Hall et al. (2008) applies unsupervised topic modeling to ACL anthology to analyze historical trends in the field of Computational Linguistics from 1978 to 206. There are not many papers out there studying the trends and dynamics between two ideas until (Tan et al., 2017). So our work attempted to further apply Chenhao's work to different fields such as Economics and different language of Mandarin Chinese.

6.2 Applications in Finance and Economics

There is growing popularity of application of Natural Language Processing in field of Economics and Finance mostly for predictive modelling and sentiment analysis based on Economics and Financial news. For example, Sesen et al. (2018) showcase different methods of getting inference from NLP models and evaluate their predictive performance based on different sources of news data. Vicari and Gaspari (2020) uses Deep Learning methods and specifically LSTM to analyze daily news headline from 2008 and 2016 for the purpose of developing optimal algorithmic trading strategy. We had very different vision: we attempt to apply idea relations analysis on the academic field of Economics with the goal of proposing social science hypothesis.

6.3 Structural Topic Modellings

Roberts et al. (2013a) develops the structural topic model that incorporate corpus structure and metadata in to the standard model and shows the model's use in two applied problems: the analysis of open-ended responses in a survey experiment about immigration policy, and understanding differing me-

dia coverage of China's rise. Since Roberts' paper, there is a growing number of application of structural topic modelling method. For example, Kuhn (2018) describes the application of structural topic modeling to Aviation Safety Reporting System data, and this application effectively reveals previously unreported issues and connections.

7 Further Exploration

Using structural topic models, we already observe interesting results on AER and People's Daily corpora, but further modifications could be made to facilitate research using the idea relations framework.

1. Different metadata. Like most real-life corpora, AER and People's Daily have numerous metadata available. For example, each People's Daily article has author, section, number of words, day in week, tone, etc. Adding these information as metadata in sTM could help us find more details in the historical ideas.

2. More textual information. With the goal to understand the change in PRC policy and ideology, we can add more text from other sources, like government announcements, other newspapers, and everyday writings of people. They could provide a more comprehensive view to the society while adding risks of mixing words with different connotations and thus making ideas illegible.

3. Combining the idea relations framework with causal inference. On the simple level, given the existing temporal structure of topics, we can locate concrete Granger causality and test them with historical knowledge or outside data. There are also other works exploring causal inference using texts (Egami et al., 2018), and it would be constructive to incorporate these ideas into ideas relations framework to allow more concrete social science knowledge to be produced.

8 Conclusion

We replicated the ideas relations framework and extend it to the more flexible structural topic model, and also explored new datasets with the framework. We discovered insights in AER and People's Daily that correspond to our existing knowledge, and used the latter to make a detailed comparison and evaluation between the two topic models. We also demonstrate that the framework has the potential to facilitate the discovery of new historical knowledge.

However, the explorations are interpretative in nature, and many of the topics and relations we discovered do not correspond to common knowledge, and it is difficult to distinguish whether they represent important trends in text and history, or they are just results of the randomness innate to our algorithm. It is also important to note that all discussion above should be viewed in the context of our corpus, which are just specific representations of the socio-historical trends we might be interested in.

References

David M Blei, John D Lafferty, et al. 2007. A correlated topic model of science. *The annals of applied statistics*, 1(1):17–35.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. page 363–371.

Kenneth D. Kuhn. 2018. Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, 87:105–122.

David M Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*, volume 24, pages 411–418. Citeseer.

Margaret Roberts, Brandon Stewart, Dustin Tingley, and Edoardo Airoldi. 2013a. The structural topic model and applied social science. *Neural Information Processing Society*.

Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. 2019. Stm: An r package for structural topic models. *Journal of Statistical Software*, 91(1):1–40.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. 2013b. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.

Alix Rule, Jean-Philippe Cointet, and Peter S. Bearman. 2015. Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, 112(35):10837–10844.

M. Berkan Sesen, Yazann Romahi, and Victor Li. 2018. Natural language processing of financial news. *Big Data and Machine Learning in Quantitative Investment*, page Chapter 10.

Matt Taddy. 2013. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Chenhao Tan, Dallas Card, and Noah A. Smith. 2017. Friendships, rivalries, and trysts: Characterizing relations between ideas in texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Mattia Vicari and Mauro Gaspari. 2020. Analysis of news sentiments using natural language processing and deep learning. *AI & SOCIETY*.