Natural Language Processing for Computational Social Science
Spring 2025
Final Project Proposal

Niyati Bafna **Krirk Nirunwiroj**
Johns Hopkins University
{nbafna1, knirunw1}@jh.edu

1 Introduction

Computation Linguistics and Natural Language Processing made fast progress over the last decade, with advances in hardware computing and GPU capacities, data availability with the rise of the internet, and neural network architectures with hugely successful Transformer-based architectures. We are interested in studying the evolution of topics of focus as well as techniques in the field over this time period.

2 Research Questions

A research paper in NLP may make an advance to the field in several ways. Its innovation may arise from novel algorithmic/mathematical approaches to problems, from linguistic insights on data, analytical insights on existing models and data, or from engineering innovations. Note that these types are orthogonal to the actual problem or subfield being studied. We are mainly interested in the following question: **How has the distribution over the above types of papers changed over time?**

We are also interested in related questions, such as how the number of papers focusing on individual fields, such as computational social science, has changed over time.

3 Implications: Why do this?

The NLP research community (like most research communities) is self-guided; i.e. there is no overarching organization who decides what the community should study, and how much effort we should invest in different things. Individual researchers and labs make decisions on relevant and impactful problems to work on based on several factors such as funding, their peers, and personal interests. Given this, it's important to present the large-picture perspective on the evolution of the field to researchers, so that we can step back and assess whether we are distributing our efforts in a reasonable manner over directions of research.

Besides, it's interesting!

4 Experimental Setup

4.1 Data

We use the ACL-OCL dataset,[1] a collection of 73,285 papers from the ACL Anthology spanning from 1952 to 2022. Each entry includes rich metadata such as title, authors, publication year, and venue (`booktitle`), along with textual content fields including `abstract` and `full_text`. A complete list of available fields is included in Appendix A.

4.2 Models

We will use neural topic models (NTMs) for unsupervised topic discovery. Our initial plan is to experiment with a selection of models implemented in PyTorch from Zhang's open-source collection[2]. These models are representative of diverse approaches to neural topic modeling and include:

- **NVDM-GSM** (Miao et al., 2017): A variational autoencoder with a Gaussian Softmax output to model topic distributions.
- **WTM-MMD** (Nan et al., 2019): A Wasserstein autoencoder with a Dirichlet prior, improving topic coherence by mitigating KL collapse.
- **ETM** (Dieng et al., 2019): Embeds both topics and words in the same space to enhance interpretability.
- **BATM** (Wang et al., 2020): Uses bidirectional adversarial training, combining GANs with encoders for robust topic inference.

[1] https://huggingface.co/datasets/WINGNUS/ACL-OCL

[2] https://github.com/zll17/Neural_Topic_Models

These models will be evaluated for feasibility given our computational constraints. If full training is not viable on our machines, we will explore large language model (LLM)-based alternatives for topic modeling via APIs (e.g., OpenAI's GPT-4, or other services that allow latent representation extraction or zero-shot topic classification).

4.3 Method and Metrics

We will first perform unsupervised topic modeling over the corpus using neural topic models (NTMs). This will produce a topic distribution for each document in the corpus, as well as a word distribution for each topic.

To categorize research papers into high-level types (e.g., engineering, analysis, modeling, linguistics), we will construct or curate small lexicons that are representative of each type. For example, the `engineering` type might include terms like `batch size`, `architecture`, and `training`, while the `linguistics` type might include terms like `syntax`, `morphology`, and `discourse`. Alternatively, if the learned topics are sufficiently interpretable, we may manually label topics based on their top keywords.

Using the topic-word distributions and these type-specific lexicons or annotations, we will assign each topic a soft distribution over research types. This could be done via lexicon overlap, cosine similarity in embedding space, or manual annotation.

We will then derive a research type distribution for each paper by aggregating over the paper's topic distribution, weighted by each topic's type distribution.

Finally, we will analyze how the prevalence of different paper types changes over time. Our primary metric is the **temporal trend** of each type's proportion. We may also report metrics such as:

- **Topic coherence:** to evaluate quality of the learned topics (using NPMI or UMass).
- **Topic diversity:** to assess the spread across vocabulary.
- **Type entropy:** to quantify how focused or mixed a paper's type distribution is.

These metrics will help evaluate both the quality of topic modeling and the interpretability of the type categorization.

Topic ID	Top Keywords
0	model, neural, models, information, propose, network, sentiment, task, based, attention
1	parsing, based, paper, task, dependency, parser, grammar, features, syntactic, tree
2	models, model, data, training, language, performance, learning, tasks, trained, domain
3	la, des, et, les, en, une, le, nous, dans, pour
4	language, corpus, annotation, data, paper, languages, research, text, present, annotated
5	semantic, paper, discourse, language, structure, linguistic, information, lexical, syntactic, analysis
6	translation, machine, english, based, mt, language, systems, question, paper, quality
7	dialogue, user, human, language, systems, users, speech, based, paper, dialog
8	task, text, evaluation, news, dataset, results, summarization, paper, information, human
9	word, words, semantic, based, method, using, approach, data, results, embeddings

Table 1: Top words from each of the 10 LDA topics trained on abstracts.

5 Summary Statistics of the Data

We restrict our analysis to papers that contain both `abstract` and `full_text` fields, resulting in a cleaned subset of **63,903** papers. This subset ensures we have sufficient content for topic modeling and text-based analysis.

We begin with basic descriptive statistics:

- **Temporal Coverage:** The cleaned data spans publications from 1952 to 2022, with the number of papers per year increasing significantly in the last two decades (see Appendix 1).
- **Venue Distribution:** The top venues include ACL, EMNLP, LREC, and COLING (Appendix 2).

We also conduct an initial LDA topic modeling analysis using 10 topics trained on the abstracts. While a few topics appear interpretable (e.g., syntax, machine translation, dialogue systems), others are vague or overlapping (e.g., multiple topics emphasize generic terms like "model" and "data"), and one consists primarily of non-English stopwords. This motivates our decision to use neural topic models for improved coherence and robustness. Full topic keywords are shown in Table 1.

These preliminary results motivate our choice to use neural topic models in the main study. They allow for more nuanced topic discovery and better semantic coherence, which is necessary for our downstream classification and longitudinal analysis.

Acknowledgments

We used OpenAI's ChatGPT to assist with grammar revision and to suggest relevant data analysis approaches for Section 5. All final decisions and writing were done by the authors.

A Data Fields

Column Name	Description
acl_id	Unique ACL ID
abstract	Abstract extracted by GROBID
full_text	Full text extracted by GROBID
corpus_paper_id	Semantic Scholar ID
pdf_hash	SHA1 hash of the PDF
numcitedby	Number of citations from Semantic Scholar
url	Link to publication
publisher	—
address	Address of conference
year	—
month	—
booktitle	—
author	List of authors
title	Title of paper
pages	—
doi	—
number	—
volume	—
journal	—
editor	—
isbn	—

Table 2: List of fields included in the ACL-OCL dataset. Dashes indicate fields with straightforward column names.

B Additional Dataset Statistics

We provide additional summary statistics of our dataset below. Figure 1 shows the number of papers per year, and Figure 2 shows the top 10 publication venues.

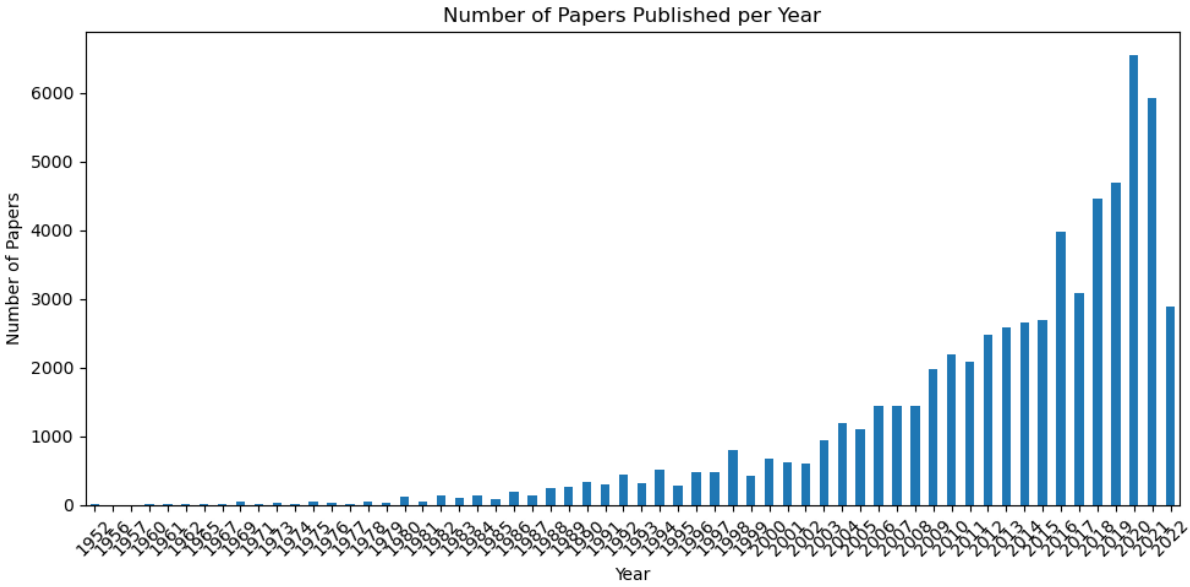

Figure 1: Number of papers per year in the filtered dataset.

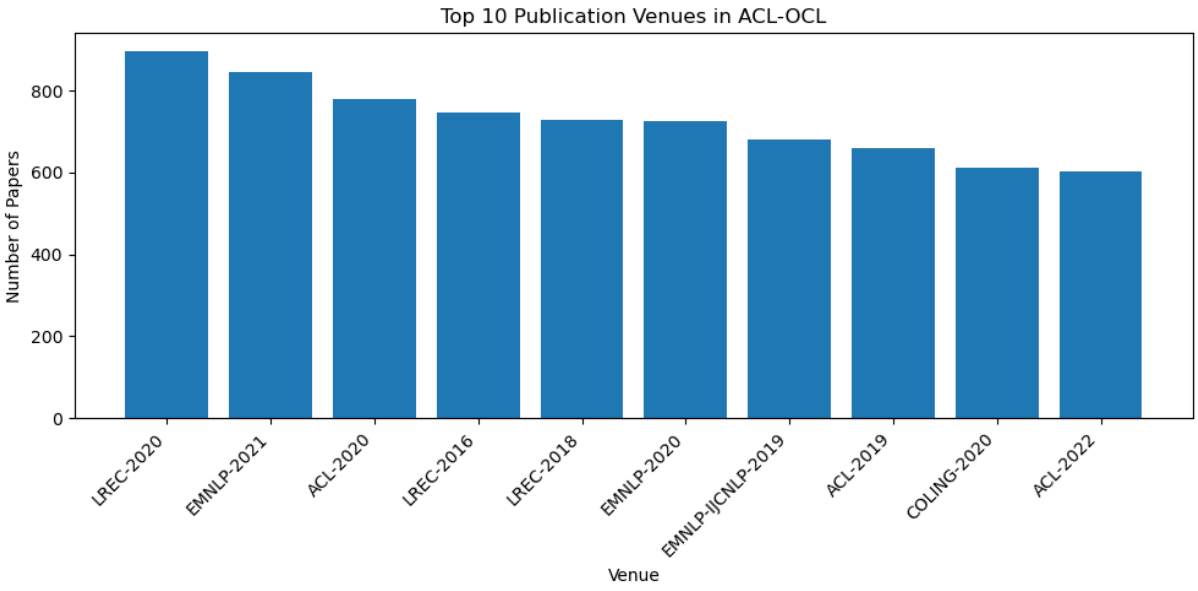

Figure 2: Top 10 publication venues by number of papers.

References

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. Topic modeling in embedding spaces. *Preprint*, arXiv:1907.04907.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2410–2419. PMLR.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381, Florence, Italy. Association for Computational Linguistics.

Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. 2020. Neural topic modeling with bidirectional adversarial training. *Preprint*, arXiv:2004.12331.