# CS 5/7322 Spring 2025

## Project

The goal of the project is to allow students to explore various problems in Natural Language Processing.

I will provide a list of projects below. Each project comes with a brief description, and some idea of what is expected. It will also come with a few papers, which you should try to read before the first meeting. Notice that some papers are available only when you logon to the SMU network (via VPN is ok).

Each project is meant to be a group project. You are to form groups of 2-3 members (please do not mix 5322 and 7322 students in the same group). Once you formed a group, you should update the group member on Canvas and e-mail me your preference of choice for project. If you cannot form a group, you should also e-mail me your preference and I will try to find group members for you.

You should rank the projects in order of preference and e-mail me back your options by 11:59pm, 3/3 (Mon). Notice this you should list all the projects in decreasing order of preference, otherwise I may put you in a project that you do not list. The project assignment will be announced 3/4 (Tue) via Canvas. I expect a maximum of 2 groups to be assigned to each project. And groups assigned to the same project may also work on different aspects of the same project.

For each project there is the following steps:

- I will meet each project group on a weekly starting 3/6 (Thu) The meeting will roughly be 15-20 minutes. I will work with each group for a time slot. For each meeting (starting the second one) there will be milestones I expect each group to finish by then. Overall progress through the milestones will count towards 25% of the project grade.
- Each 7330 student (including those in distance section) will need to present a paper related to the project between the class of 4/21 – 4/30. Those who are in distance section will record your presentation and upload it to Canvas (or provide a YouTube link). A further announcement will be made about the schedule.
- Each group will need to present their work on 5/8 (Thu) between 3-6pm. Each group will have around 10-12 minutes for its presentation. More details will be provided later. This will count towards 15% of the grade
- The final deliverables for each project need to be uploaded to Canvas (as a zip file) by 11:59pm 5/12 (Mon). This will count towards 60% of the grade.

You are also welcome to propose your own projects. The list of projects here gives you a rough guideline of the type of projects that I am interested in.

**List of projects**

1. Effect of corpus for language model performance

   In this project we would like to train various language models (e.g. BERT, T5 etc) using different corpus – for example, two corpus that are the same except for the document for some keywords removed. We would like to see what effect the difference in corpra does have on downstream tasks.

Papers:

- Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, and Nako Sung. 2022. On the Effect of Pretraining Corpora on In-context Learning by a Large-scale Language Model. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5168–5186, Seattle, United States. Association for Computational Linguistics.

- Chen, Hao, et al. "On the Diversity of Synthetic Data and its Impact on Training Large Language Models." *arXiv preprint arXiv:2410.15226* (2024).

- Zhao, Yang, et al. "Deciphering the impact of pretraining data on large language models through machine unlearning." *arXiv preprint arXiv:2402.11537* (2024).

2. Study on prompt-based learning using OpenPrompt

   Prompt learning is one method that is trying to leverage the basis of large language model (LLM) to improve the usage of LLM for downstream NLP tasks. This project uses the OpenPrompt (and/or other) framework for prompt learning to test the performance of prompt-learning task

   Papers:

   - Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Comput. Surv. 55, 9, Article 195 (September 2023), 35 pages. https://doi.org/10.1145/3560815 (2 students combined to present the paper: total 25-30 minutes)

   - Ding, Ning, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. "Openprompt: An open-source framework for prompt-learning." *arXiv preprint arXiv:2111.01998* (2021).

   Software:

   - GitHub - thunlp/OpenPrompt: An Open-Source Framework for Prompt-Learning.
   - GitHub - microsoft/promptbench: A unified evaluation framework for large language models (This will be used by other projects)

3. Incremental learning

One challenge about Large Language Models is that they are built from a fixed corpus. So, in real life there is a challenge in how to incorporate new knowledge into the model. In this project we want to experiment with some ways of doing that, and measure its performance (especially with conflicting data)

Papers:

- Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997* 2 (2023). (2 student combine to present the paper: total 25-30 minutes)
- Du, Mingzhe, et al. "From static to dynamic: A continual learning framework for large language models." arXiv preprint arXiv:2310.14248 (2023).

4. Handling conflicting knowledge in large language models

Given that a large corpus is used to train a large language model, it is unavoidable that it will contain conflicting knowledge. In this project we will study how this kind of situation is discovered and managed and will implement some methods that will handle these situations.

Papers:

- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge Conflicts for LLMs: A Survey. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.
- Hou, Yufang, et al. "WikiContradict: A Benchmark for Evaluating LLMs on Real-World Knowledge Conflicts from Wikipedia." *Advances in Neural Information Processing Systems* 37 (2025): 109701-109747.
- Xie, Jian, et al. "Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts." *The Twelfth International Conference on Learning Representations*. 2023.

Data

- ibm-research/Wikipedia_contradict_benchmark · Datasets at Hugging Face

5. A tool to classify and tag political leaning in articles

There has been quite a bit of work done on text classification based on political inclinations (e.g. conservative vs. liberal). In this project you would use some established method, together with some tools you build, to tag sentence/articles/news sources based on political inclination.

Papers:

- Di Giovanni, Marco, et al. "Content-based classification of political inclinations of twitter users." *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018.

- Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.
- Kornraphop Kawintiranon and Lisa Singh. 2022. PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7360–7367, Marseille, France. European Language Resources Association.

Data sets:

- newsmediabias/news-bias-full-data · Datasets at Hugging Face
- GitHub - irgroup/Qbias: $Qbias$ - A Dataset on Media Bias in Search Queries and Query Suggestions
- GitHub - ramybaly/Article-Bias-Prediction

6. Developing a FrameNet parser

   The goal for this project is to apply various machine learning techniques (tagging, neural networks etc.) to develop a tool that determine the frames that are relevant for a given sentence, and discover the frame elements.

   Papers:

- Baker, Collin F., Charles J. Fillmore, and Beau Cronin. "The structure of the FrameNet database." *International Journal of Lexicography* 16.3 (2003): 281-296.929–936, Sydney, Australia.
- Das, Dipanjan, et al. "Frame-semantic parsing." *Computational linguistics* 40.1 (2014): 9-56.
- Kalyanpur, Aditya, et al. "Open-domain frame semantic parsing using transformers." *arXiv preprint arXiv:2010.10998* (2020).

Data sets/tools:

- Welcome to FrameNet! | fndrupal
- NLTK :: Sample usage for framenet
- GitHub - machinereading/frameBERT

7. Pronoun resolution as a preprocessor for Topic Modelling (and other tasks)

   Pronoun resolution has been studied extensively in NLP. While this is a standalone task on its own right, one can also use it to pre-process a text document for downstream task. For this

project I would like to explore applying pronoun resolution as a pre-processing task because applying the documents to some other NLP task (with a focus on topic modeling).

Papers:

- Rhea Sukthanker, Soujanya Poria, Erik Cambria, Ramkumar Thirunavukarasu Anaphora and coreference resolution: A review, Information Fusion, Volume 59, 2020, Pages 139-162, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2020.01.010.
- Rakesh Chada. 2019. Gendered Pronoun Resolution using BERT and an Extractive Question Answering Formulation. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 126–133, Florence, Italy. Association for Computational Linguistics.

Data sets & Tools:

- community-datasets/definite_pronoun_resolution · Datasets at Hugging Face
- GitHub - HKUST-KnowComp/PCR4ALL: This is the github repo for LREC 2022 paper "PCR4ALL: A Comprehensive Evaluation Benchmark for Pronoun Coreference Resolution in English".
- GitHub - Yorko/gender-unbiased_BERT-based_pronoun_resolution: Source code for the ACL workshop paper and Kaggle competition by Google AI team
- GitHub - huggingface/neuralcoref: ✨ Fast Coreference Resolution in spaCy with Neural Networks

8. SQL to text

There has been a lot of work on generating SQL queries from natural text. However, relative few work has been done on the reverse process: given an SQL query, together with a textual description of the fields and tables, convert the SQL query into text. In this project we would like to explore a set of techniques that try to solve the problem, and implement a couple of those.

Papers:

- G. Koutrika, A. Simitsis and Y. E. Ioannidis, "Explaining structured queries in natural language," *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, Long Beach, CA, USA, 2010, pp. 333-344,
- Kun Xu, Lingfei Wu, Zhiguo Wang, Yansong Feng, and Vadim Sheinin. 2018. SQL-to-Text Generation with Graph-to-Sequence Model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 931–936, Brussels, Belgium. Association for Computational Linguistics.
- V. Câmara, R. Mendonca-Neto, A. Silva and L. Cordovil, "A Large Language Model approach to SQL-to-Text Generation," *2024 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, 2024, pp. 1-4, doi: 10.1109/ICCE59016.2024.10444148.

Data and tools:

- Spider: Yale Semantic Parsing and Text-to-SQL Challenge (for the data only)
- GitHub - andialbrecht/sqlparse at hackernoon.com

9. Stance detection tool.

   Different news outlets can choose to cover different news events with different or take a different stance. In this project we want to use various NLP tools to discover and summarize stance on various articles on a topic.

   Papers:

   - Alturayeif, N., Luqman, H. & Ahmed, M. A systematic review of machine learning techniques for stance detection and its applications. *Neural Comput & Applic* **35**, 5113–5144 (2023).
   - Mayor, E., Miani, A. A topic models analysis of the news coverage of the Omicron variant in the United Kingdom press. *BMC Public Health* **23**, 1509 (2023).
   - Gül, Ilker, Rémi Lebret, and Karl Aberer. "Stance detection on social media with fine-tuned large language models." *arXiv preprint arXiv:2404.12171* (2024).

   Data set/Tools:

   - GitHub - Ayushk4/stance-dataset: The Pytorch code for baselines and dataset accompanying the NAACL 2021 paper - "tWT–WT: A Dataset to Assert the Role of Target Entities for Detecting Stance of Tweets"