CS 5/7322 Spring 2024

Project

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

I will provide a list of projects below. Each project will comes with a brief description, and some idea of what is expected. It will also come with a few papers, in which you should try to read before the first meeting. You should rank the projects in order of preference and e-mail me back your options by noon, 3/4 (Mon). The project assignment will be announced on Monday's class. 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 (Wed) The meeting will roughly be 10-15 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 group will need to present its work on 5/6 (Mon) 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 5/7 (Tue). This will count towards 60% of the grade.

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

List of projects

1. "Forgetting"/"Changing" language models

   We want to explore how will a pre-trained changed if one feed in more data to train. Especially if the newly trained data is very biased and opposite to what have been learned. We will try to see how easy (or hard) such models can be affected.

   Papers:

   - Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, Quoc Viet Hung Nguyen, A Survey of Machine Unlearning. https://doi.org/10.48550/arXiv.2209.02299
   - Ronen Eldan, Mark Russinovich. Who's Harry Potter? Approximate Unlearning in LLMs, DOI:10.48550/arXiv.2310.02238

2. Overcoming LDA instability

One challenge of using topic models like LDA is the instability of it. For instance, running LDA on the same set of documents for multiple times likely will return different topics. This project explore various methods that will enable a coherent results of topics to be output.

Papers:

- Yi Yang, Shimei Pan, Jie Lu, Mercan Topkara, and Yangqiu Song. 2016. The Stability and Usability of Statistical Topic Models. ACM Trans. Interact. Intell. Syst. 6, 2, Article 14 (August 2016), 23 pages. https://doi.org/10.1145/2954002
- Yi Yang, Shimei Pan, Yangqiu Song, Jie Lu, and Mercan Topkara. 2016. Improving topic model stability for effective document exploration. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16). AAAI Press, 4223–4227.
- Mika V. Montyla, Maelick Claes and Umar Faroaq, Measuring LDA topic stability from clusters of replicated runs, *In Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '18)*, 2018.

3. A library of short text analysis tool.

Many text algorithm (topic modelling, classification etc.) may work different if the input is a set of short text (e.g. social media posts). In this project I would like to be a toolset (leveraging existing library like genism, nltk, and spacy) that allow processing of short text efficiently.

Papers:

- Qiang, Jipeng, et al. "Short text topic modeling techniques, applications, and performance: a survey." *IEEE Transactions on Knowledge and Data Engineering* 34.3 (2020): 1427-1445.
- Alsmadi, I., Hoon, G.K. Term weighting scheme for short-text classification: Twitter corpuses. *Neural Comput & Applic* **31**, 3819–3831 (2019). https://doi.org/10.1007/s00521-017-3298-8
- Ji Young Lee and Franck Dernoncourt. 2016. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520, San Diego, California. Association for Computational Linguistics.

4. An analysis of wishes

People have been making wishes, expressing needs and desire and so on. In this project we want to analyze wishes and/or desires that people express and try to categorize them and develop tools to process them specifically.

Papers:

- Andrew B. Goldberg, Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson, and Xiaojin Zhu. 2009. May All Your Wishes Come True: A Study of Wishes and How to Recognize Them. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 263–271, Boulder, Colorado.
- Kühl, Niklas, and Gerhard Satzger. "Needmining: Designing digital support to elicit needs from social media." *arXiv preprint arXiv:2101.06146* (2021).Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020.
- Ao Jia, Yu He, Yazhou Zhang, Sagar Uprety, Dawei Song, and Christina Lioma. 2022. Beyond Emotion: A Multi-Modal Dataset for Human Desire Understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1512–1522, Seattle, United States. Association for Computational Linguistics.

5. 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 relavent 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).

6. Did pronoun (anaphora) resolution help in other NLP task?

   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 apply pronoun resolution as a pre-processing task because applying the documents to some other NLP task (e.g. topic modelling).

   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.