Kate Salembier
Annotated Bibliography
Natural Language Processing Reading Group

Ananya Ganesh, Jie Cao, E. Margaret Perkoff, Rosy Southwell, Martha Palmer, Katharina Kann. 2023. "Mind the Gap between the Application Track and the Real World"

As NLP advances, more studies are becoming application-oriented. This means that research increasingly utilizes experiments that do not account for the nuance of real-world data or situations. This position paper investigates the relationship between the stated applications of the research and the experimental tasks used to test the models. The authors conducted a survey of the papers from ACL 2020 and EMNLP 2020. They had three authors evaluate the papers on the following questions: Does the paper comprehensively describe the use case for a reader to understand? Is the paper dealing with an entire task or a sub-task only? Does the paper mention the other missing subtasks explicitly? Is the downstream evaluation realistic? They found that many of the papers utilized simplified evaluation methods that did not consider real-world circumstances and applications. Additionally, they conducted a case study that compared the performance of an educational dialogue. They found that the dialogue deteriorated when used in an actual educational setting. They acknowledge in their results that it is a long-standing tradition to contextualize NLP research within the potential applications of the task; however, they urge readers to more critically analyze how experimental methods may or may not reflect real-world applications. I think this was a very interesting paper and an apt choice to close out the semester. I think that they make strong points about the healthy dose of apprehension we should approach the relationship between experimental methods and proposed applications. I will definitely be keeping this paper in mind as I continue to read more journals in the future.

Anthony Sicilia, Jennifer C. Gates, Malihe Alikhani. 2024. "HumBEL: A Human-in-the-Loop Approach for Evaluating Demographic Factors of Language Models in Human-Machine Conversations"

This paper suggests that there is a gap in the research on how pre-trained LMs can adapt to the differences in communication that naturally arise across demographics like age and gender. To address this, the authors consider how they might be able to measure demographic factors and determine compatibility with various target demographics. In order to explore this, they suggest techniques from clinical speech pathology practices that have baselines for language acquisition skills. They conducted evaluations with a clinically licensed speech pathologist and also proposed various automated techniques. They then introduce HumBEL which is an experimental framework that evaluates demographic factors in LMs by using "novel clinician-in-the-loop statistical techniques." They find that while HumBEL has been applied to communication gaps between LMs and humans in this study, there might be other applications for this framework. Language disorders, simulated worlds, and cross-cultural human-machine connections are all

examples of potential uses for this framework. I found this to be a very interesting study and appreciate that the framework that was attempting to understand humans engaged with humans. I would be interested to see how this framework can be applied to other scenarios.

Jacob Mitchell Springe, Suhas Kotha, Daniel Fried, Graham Neubig, Aditi Raghunathan. 2024. "Repetition Improves Language Model Embeddings"

This paper identifies and attempts to provide a solution to a structural limitation of autoregressive LLMs. Autoregressive LLMs are models that predict the next token in a sequence given the previous tokens. A side effect is that "token embeddings cannot contain information from tokens that appear later in the input." As a result, the authors propose "echo embeddings," or embeddings that repeat the input and extract embeddings from the second set of data. In order to test this method, they compared classical and echo embeddings in a zero-shot setting and with fine-tuning. They tested these models with toy data and discovered that the echo embeddings were able to recover where classical embeddings failed. I found this paper very challenging on the first read, but after a significant amount of time spent googling various terms, I felt more confident in my understanding. I think one of the most impressive things about this article is the relatively simple solution to a problem that has existed for all autoregressive LLMs to date. I think this definitely expanded my understanding of various benchmarks and embedding strategies.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, Daphne Ippolito. 2023. "A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity"

This paper emphasizes the importance of pretraining data design and explores the impact of toxicity and quality filters. When developing LLMs, pretraining is the preliminary stage, where the model is trained on curated data. This article reports that preraining data design is "critically under-documented" and attempts to address this by pretraining 28 models under various conditions. They trained the models on data curated at "different times, with varying toxicity and quality filters, and with different domain compositions." In regards to models trained at different times, they found that the temporal distance between pretraining data and evaluation data leads to a decreased performance that can not be remedied with fine-tuning. In regards to toxicity and quality filters, they found that there is a trade-off between increased performance and the risk of toxic generations. They suggest that there is no general rule that can be applied but that all pretraining and different types of filtering are not predictable across various data sets. Finally, in regard to model training with different domain compositions, they found that the inclusion of heterogeneous data sources is widely beneficial and should be prioritized in future research. The authors suggest that these findings are the largest set of experiments on pretraining conditions and "expose many undocumented institutions." It seemed to me that most of the points they made about pretraining and the necessity of documentation are

common sense, but clearly, not enough institutions are prioritizing this, so I can only hope this paper informs pretraining standards in the future.

Kent K Chang, Mackenzie Cramer, Sandeep Soni, David Bamman. 2023. "Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4."

These authors were interested in exploring which books are known to ChatGPT and GPT-4 and subsequently deployed a *name cloze* membership inference query. This method involves "probing the degree of exact memorization for a sample of passages from 571 works of fiction." This task resulted in various findings. The first is that OpenAI models have memorized a significant amount of copyright books. The second is that there is a systematic bias in what books OpenAI has memorized, and the most strongly memorized are popular works in the public domain. Next, they found that this bias aligns with general web activity and confirms that repeated appearance on the web encourages memorization. Finally, they report that "disparity in memorization leads to disparity in downstream tasks."

Overall, this paper provides very interesting information about the kind of data that ChatGPT and GPT-4 were trained on and raises some concerns over copyright infringement and intellectual property. Some additional pitfalls or loopholes discussed by the group are questions about the LM's ability to detect compound names or names with prefixes (Dr., Mrs., etc.). We also discussed the implications of fan fiction on this system as it would contain characters and themes from published work but would differ from the source text. It would be interesting to investigate how this impacts the LM.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, Yejin Choi. 2020. "The Curious Case of Neural Text Degeneration."

This article explores the question of the best decoding method for text generation from a language model. They suggest that maximization-based decoding leads to text degeneration and proposed Nucleus Sampling as an alternative. Nucleus Sampling avoids degeneration by "truncating the unreliable trail of the probability distribution, sampling from the dynamic nucleus of tokens containing the vast majority of the probability mass." My understanding is that, rather than only taking into account the previous word, all of the tokens are considered when generating the next word in the sentence. Their results confirmed that maximization is not the best method for open-ended text generation, that current LLMs would benefit from truncating their unreliable trail, and that Nucleus Sampling is best for generating long-form text. I definitely found this paper challenging to understand. I think our discussion helped me understand some of the research methods, but overall, this paper was very technical, and I am not confident in my understanding of the topic. However, my conclusions are that this method would reduce bland, repetitive, or incoherent output, which would be a helpful development in the field of NLP.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, Aida Nematzadeh. 2023. "Pragmatics in Language Grounding: Phenomena, Tasks, and Modeling Approaches."

This position paper emphasizes the importance of context to successful communication, analyzes current pragmatic modeling tasks, and provides recommendations for future task design. This article opens by exploring the impact of context on language and then introduces the subject of pragmatics and types of pragmatic phenomena. Next, the authors discuss the existing tasks used to evaluate pragmatic skills in LLMs. These tasks included reference games, image captioning, instruction following, and grounded goal-oriented dialogue. Finally, this paper made suggestions for the direction of future research. The authors encouraged future investigation to "focus on realistic interactive scenarios" and contextualization of existing NLP tasks in order to increase "real-world applicability." They also suggest that future work should improve human evaluations and focus on engaging and motivating tasks. This paper made strong observations about the state of current pragmatic modeling and salient suggestions for future research. I think if the recommendations of this paper are applied to future studies, it will advance NLP to previously unknown capabilities.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Mohit Iyyer. 2023. "TopicGPT: A Prompt-based Topic Modeling Framework."

This article introduces TopicGPT, which is a framework for topic modeling that performs better than various other topic modeling approaches. Rather than just grouping topics as "bags of words," TopicGPT uses a "prompt-based framework" that produces topics that more closely resemble human-generated categorizations. Additionally, it was reported that TopicGPT produced higher quality and more interpretable topics than competing approaches. By using natural language labels and descriptions, TopicGPT allows users to bypass an additional labeling step. TopicGPT was also found to be more customizable as it allows users to input an initial set of example topics and then manually make further adjustments once TopicGPT has generated topics. Furthermore, while other topic generators are competent at assigning topics, ChatGPT has proven to be better at generating topics and identifying topics that are more often aligned with human-annotated labels. When TopicGPT did not align with ground truth labels, the labels were still found to be semantically similar, and when further prompted to provide multiple labels, including the ground truth label as well as the original label. This article provided information regarding topic modeling, a subject I was previously unfamiliar with, and outlined various frameworks and baselines under which they were evaluating TopicGPT. This paper expands our understanding of LLMs and how they can be leveraged to illuminate latent topics in text corpora.

Naitian Zhou, David Jurgens, David Bamman. 2023. "Social Meme-ing: Measuring Linguistic Variation in Memes."

This paper investigates memes as a multimodal form of language using a data set of 3.8 million memes collected from Reddit communities and finds that "patterns of meme innovation and acculturation within these communities align with previous findings on written language." I think this study was well conducted and contributed meaningful findings to an emerging field of study. I think there are a few limitations to this study, and the methodology could be clarified for

non-experts, but overall, I felt it was an engaging article and furthered my understanding of analysis models for image, text, and a combination of both. One of the findings is that in a subreddit, community members will take publicly popular memes and create semantically similar memes that are more specifically related to the subject of their community. It was also observed that the longer one is in a subreddit, the more likely it is that they will use memes associated with that community. During our discussion, it was brought up that this does not account for the fact that the oldest members of any given community are likely the ones originating the new memes. Another possible limitation of the study was overlooking the impact of moderators. Sometimes, moderators limit or influence memes, which could bias the data set. I think the future of this research could investigate just text-based memes and just image-based memes or reaction images.

Sagi Shaier, Lawrence E. Hunter, Katharina von der Wense. 2023. "Who Are All The Stochastic Parrots Imitating? They Should Tell Us!"

This position paper makes an argument that LMs "their current state will never be fully trustworthy," and suggests that they cite their sources as a novel strategy to solve this problem. They go on to suggest that preliminary data cleaning might be utilized to determine that all retrieved knowledge is factual and from strong sources. This brought up questions regarding who was doing the data cleaning, humans or AI, and furthermore, what criteria will be used to decide if information is a "good source"? It was suggested that this would likely introduce at least some bias into the data set. This seemed a particular problem when applied to discussions of low-resource LMs. The article then discusses the pros and cons of citations. The primary benefits outlined were the ability to verify information, prevent copyright violations, allow users to use citations for further research, and improve the trustworthiness of AI. The main negatives mentioned were that citations would improve trust in AI but that users would likely not fact-check every source. Additionally, readability may be reduced, and sensitive or private information may be revealed. I found this article interesting, and the idea of citations is a good one. My main concern with their proposed model was the idea of filtering data and how it might be accomplished. I think any criteria applied would have to be rigorously vetted. Additionally, I found it concerning that citations might create a false sense of security for users and might perpetuate issues of blind trust in AI. As it is, people are appropriately suspicious of information provided by LMs and will fact-check if they deem it necessary. The addition of citations might increase misinformation by inadvertently discouraging users from verifying information.