LLM discussion

Ann: What do we discuss today vs. other related discussions?
Dan: Proposed SIG for distinguishing human-authored vs machine-generated text
Ann: Three possible points of connection of D and LLMs: 1) using D resources to distinguish between human-authored output and LLM output. Based on brief experiments: there is parse-level diversity metrics where this seems to work well. Frequency of constructions. There is work on vocabulary ("delve") but there is differences in syntax. 2) Synthetic data is of huge interest now. Kuhnle (Shapeworld). Interesting results that were then used. Kuhnle's thesis is important for people who are interested in taking NL evaluation from fixed (data?)sets. 3) ?

Dan: On the first issue, in the experiments on the Cambridge Grammar of English and frequencies of constructions there, I was looking at 15K examples with a very long tail. That tail curve may be a footprint of the language use as recorded in CGE. I want to take a block of text produced by LLM, parse it with top-1 ERG mode, count constructions there, look at the footprint, compare it to perhaps the footprint of Wikipedia. Size: maybe 100K words. I expect the shapes to be dramatically different. Will need to repeat that and maybe see patterns in the differences of the shapes of these long tails. Example: Inverted sentence conditional: "were we to read this text carefully, we would understand it better". It does show up un human texts quite a bit. Maybe not in machine-generated text? (That's a hypothesis.) The idea is to have a tool that can be applied in general.

Ann: Re using LLM in academic (student) work: we don't have enough data to have a robust procedure. There are plenty of humans who use "delve" a lot (probably whoever authored that training data). So you may be biasing the tool against people e.g. from countries that generated the training data.

Emily: Agreed. As soon as we put something out there, it is going to be used to police e.g. students.

Dan: So may this shouldn't be published but the puzzle is still interesting

Emily: What about not looking at individual texts but looking at texts before and after LLM was released. No framing: "can I tell for a particular text"

Ann: Also would have to look at differences at specific points still. Due to randomness in parse ranking, what might work out better is clustering (??) Comparing different metrics would be better. There are other applications such as determining the provenance of the training data (e.g. lots of web-pages that are autogenerated that you don't necessarily want in your training data).

Emily: Also news sources

Dan: The website task doesn't have the same concerns with bias?

Emily: It does if you can use the tool for the same purpose of policing students via the API

Luis: ??

Emily: Can use a big minimal size (something a student won't produce)

Ann: Curating data: excluding some human-authored text is less problematic

Emily: if you do it systematically, still have the same bias concern  with the end use

Ann: if we find interesting pockets of data, they can lead us to creating better synthetic data

Ann: Synthetic data. LLM training will run out of data within the next couple of years. Soon all the data will be automatically generated. Moral issue: energy, climate crisis. Grounded synthetic data. Visual genome, semi-grounded. More efficient for BERT-style models.

Point 3): Training "baby LLMs". Not just synthetic data, but semi-grounded parse data could be very useful. Some identification of the entities involved. We could use MRS semantics there.

Alex L.: Shapeworld indeed useful for 0-shot learning and connecting predicate symbols to visual features. A student developed an interactive model. You can develop architecture for embodied situations where LLMs get things wrong. But I don't agree that D resources will solve the synthetic data problem because LLMs need novels, stories. Principled manner of organization. And that's what LLMs are after. The major advantage is LLMs are very context sensitive, and grammars are not. You need a computational model of pragmatics. We don't have it.

Ann: Agreed. Still, for now there is no alternative good source of synthetic data.

Alex L.: Interactive task learning. Human interacting with an autonomous system that adapts the environment. Domain expert to teach the model, then 0-shot learning.

Ann: We can provide types of synthetic data that people haven't thought of.

Alex L: No current vision language model that can distinguish species of apples, and we used the ERG to teach it quickly to do it (reference?) Demonstrative case of F1 increase from before interactive learning with Delphin to after

Emily: all test data is in training (probably). So, evaluation maybe is the domain

Guy: publishing non-LLM datasets is easier than non-LLM systems.

Alex L: as for the long-tail, you need to find where that long-tail causes problems to LLMs

Ann: Ethics concerns. As for evaluation, Kuhnle's thesis (1st part again).
Ann: Training/test data for human-satisfactory explanations (as opposed to logical explanations). We can find publishing venues. But where to get the money from?

Luis: At VU we have a group interested in baby LMs. Explainability. Hack the learning by ordering the data and playing with structure types. Some baby LMs can perform at 70% to LLAMA-2. Not all sentences are the same. We can tell how they are different. Bootstrapping: we can filter the language the LMs produce before it is fed back to them. At which layers certain types of info is encoded? Also need to explore multilingual models (in the absence of massive amounts of data).

Eric: two places of intersection of LMs and D, which I explored: (1) in some cases, I used LLM as an oracle in very constrained situations. Robustness (if you hit an UNK). (2) generation I have is poor but I can use LLM to generate, but I want to use D to get a "percent chance" that the output is off the rails. Comparison of the texts that something weird has got in. Doesn't need to be perfect, but to increase confidence that I can use LLM output.

Dan: Another use case: use LLM as a smoothing of slightly ungrammatical input. Very effective.

Ann: Paraphrase-type applications fit well with LLMs. Increase applicability of small models.

Eric: Paraphrasing or simplifying

Emily: with using LMs to do paraphrasing, how safe are we around e.g. negation? Do I see negation in one but not another version?

Ann: Negation is way less of a problem now for some language pairs in MT than it used to be

Olga: using LMs to have something like a github copilot

Mike: that's not LMs, that's more like a language-specific IDE

Luis: Perhaps more like adding lexical entries. And of course fine-tuning.

Ann: You might not need *that* much data, there have been experiments on something similar. We could tray training models for MRS2LEAN (?) conversion?

Emily: would it be harder to verify the output of autogenerated grammar code than to have written it by hand in the first place

Francis: Students say translation is worse between e.g. Czech-Korean than Czech-English or Korean-English. Assuming we have all the grammars

Olga: haha

Francis: We could produce data for parallel translation sets.

Guy, Francis: discussing honorifics as a phenomenon that is important for Czech and Korean