Inference in the Time of GPT*

Mark Steedman *with* Nick McKenna, Tianyi Li, Liang Cheng,
Javad Hosseini, Liane Guillou, and others

* and apologies to G. Garcia Marquez

Mar 21st 2023

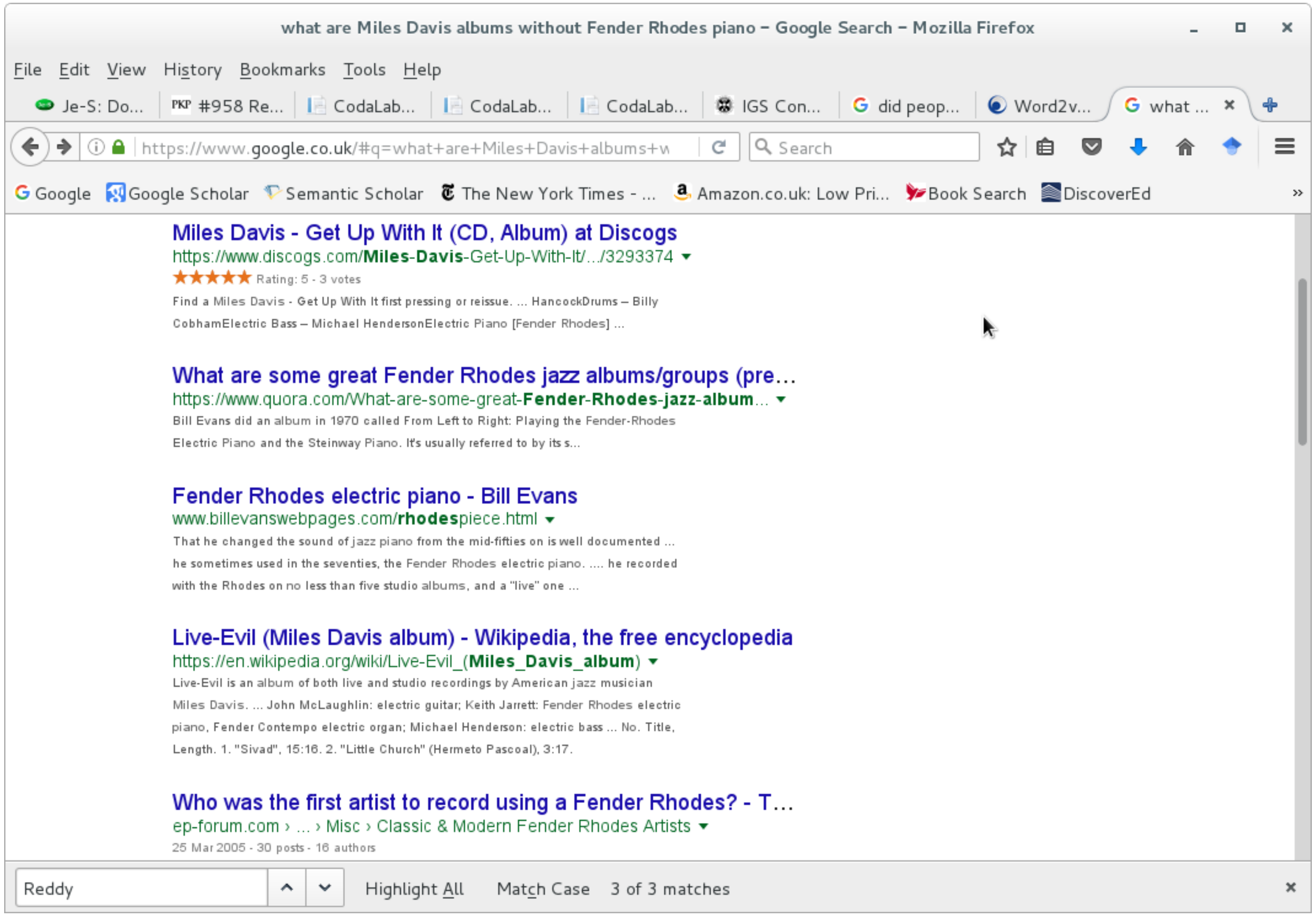
School of informatics

The State of the Art in Open-Domain QA

- Where are we? Magical thinking about GPT-3.
- Where should we be? Combining Logic- and Language-model -based methods (Fan, Gardent, Braud, & Bordes, 2019.)
- How do we get there?
  - By making semantics scale using Meaning Postulates;
  - By exploiting the way language models actually represent text (rather than believing they must somehow be learning latent syntax and semantics, and even inference).

The Problem of Open Domain QA

- There are too many ways of asking and answering the same question:
- You want to know Who played against Manchester United? The text says:
  - Arsenal beat Manchester United.
  - Manchester United's defeat by Arsenal.
  - Arsenal obliterated Manchester United.
  - etc.
- So if you just build a knowledge graph based on relations found in text (a "Semantic Network"), you won't be able to interrogate it effectively.
- Why not just Google it!
  - "What are Miles Davis records without Fender-Rhodes piano?"


what are Miles Davis albums without Fender Rhodes piano – Google Search – Mozilla Firefox
File Edit View History Bookmarks Tools Help
Je-S: Do... #958 Re... CodaLab... CodaLab... CodaLab... IGS Con... did peop... Word2v... what ...
https://www.google.co.uk/#q=what+are+Miles+Davis+albums+w
Search
Google Google Scholar Semantic Scholar The New York Times - ... Amazon.co.uk: Low Pri... Book Search DiscoverEd
Miles Davis - Get Up With It (CD, Album) at Discogs
https://www.discogs.com/Miles-Davis-Get-Up-With-It/.../3293374
Rating: 5 - 3 votes
Find a Miles Davis - Get Up With It first pressing or reissue. ... HancockDrums – Billy CobhamElectric Bass – Michael HendersonElectric Piano [Fender Rhodes] ...
What are some great Fender Rhodes jazz albums/groups (pre...
https://www.quora.com/What-are-some-great-Fender-Rhodes-jazz-album...
Bill Evans did an album in 1970 called From Left to Right: Playing the Fender-Rhodes Electric Piano and the Steinway Piano. It's usually referred to by its s...
Fender Rhodes electric piano - Bill Evans
www.billevanswebpages.com/rhodespiece.html
That he changed the sound of jazz piano from the mid-fifties on is well documented ... he sometimes used in the seventies, the Fender Rhodes electric piano. .... he recorded with the Rhodes on no less than five studio albums, and a "live" one ...
Live-Evil (Miles Davis album) - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Live-Evil_(Miles_Davis_album)
Live-Evil is an album of both live and studio recordings by American jazz musician Miles Davis. ... John McLaughlin: electric guitar; Keith Jarrett: Fender Rhodes electric piano, Fender Contempo electric organ; Michael Henderson: electric bass ... No. Title, Length. 1. "Sivad", 15:16. 2. "Little Church" (Hermeto Pascoal), 3:17.
Who was the first artist to record using a Fender Rhodes? - T...
ep-forum.com › ... › Misc › Classic & Modern Fender Rhodes Artists
25 Mar 2005 - 30 posts - 16 authors
Reddy
Highlight All Match Case 3 of 3 matches

Open Domain QA Needs Inference

- Webber, Gardent, & Bos, 2002 give more QA examples, including
  - Query expansion to entailing alternatives;
  - Eliminating spurious answers;
  - Eliminating redundant alternative answers;
  - Detecting equivalence to FAQs;
  - Generating explanatory answers.
- Fan, Gardent, Braud, & Bordes, 2019 Multi-Document summarization:
  - "General Relativity is a theory of Albert Einstein. Einstein developed this theory."
- These are tasks where precision matters!

The Problem of Inference

- The problem arises from the lack of a usable NL semantics supporting common-sense inference, such as that $\langle team\rangle defeat\langle team\rangle$ entails $\langle team\rangle play\,against\langle team\rangle$, $\langle recording\rangle without\langle musical\ instrument\rangle$ entails $\langle recording\rangle \wedge \neg with\langle musical\ instrument\rangle$, and $\langle theory\rangle of\langle person\rangle$ entails $\langle person\rangle develop\langle theory\rangle$.

- Two solutions:

  1. Use of a pretrained LM, such as BERT or GPT-3, as a latent entailment model, with or without Supervised fine-tuning using an NLI dataset, "Train-of-thought" prompting, "In-context learning", etc.;
  2. Unsupervised induction of an entailment graph from text, using some form of the Distributional Inclusion Assumption (Geffet and Dagan, 2005).

LMs as Latent Entailment Models

- Schmitt and Schütze (2021b,a) claim that fine-tuning BERT/RoBERTa LM using NLI training datasets makes it learn entailment, as assessed on NLI test-sets.
- They embedded entailment pairs in text-like patterns, such as "P, and so Q".
- However, evaluating supervised text inference is an open problem: NLI datasets are:
  - Riddled with artefacts that ML can learn as a proxy;
  - Dominated by paraphrase and selection-bias; and
  - Fail to include false inverses of directional entailments.
- When these artefacts are properly controlled for, Li *et al.* (2022a) fail to support Schmitt and Schütze's claims.
- RoBERTA seems to model mere non-directional associative similarity.

Very Large LMs as Entailment Models

- Some of our current work investigates Very Large Language Models such as GPT-3 as entailment models.
- VLLMs appear to memorize the training data, and to organize the memory according to similarity of textual context.
- The larger they are, the more literally this is the case (Zhang *et al.*, 2021; Tirumala *et al.*, 2022)
- They excel at tasks where the memorized text actually contains something similar to the question (particularly with respect to nouns and named-entities).
- GPT-3.5 has been tweaked by fine-tuning on all kinds of task-oriented data, probably including NLI datasets.

VLLMs as Entailment Models

- We therefore embed entailment test pairs in MNLI-like Schmitt and Schütze prompts: eg.:

  "If Google bought YouTube, then Google owns YouTube.

  A) Entailment

  B) Neutral

  C) Contradiction

  Answer:"

- When we test Zero-shot with these patterns, GPT-3 does quite poorly:

Pattern GPT-3.5	Precision	Recall	F1
With Named Entities:	53.4	79.7	64.0
With Entity Types:	53.1	52.9	54.0
With Untyped ABC:	53.03	44.0	48.1
All-positive baseline	50.0	100.0	**66.7**

VLLMs as Entailment Models

- Two-shot "Train-of-Thought" prompt training with a pair of such examples as prefix augmented with an explanation for the decision ("owning is a consequence of buying") prefixed to each MNLI-syle text item adapted from Levy Holt Directional Subset got F1 of **74.3** with full named entities.
- It was still quite bad at rejecting non-entailing inverses.
- —consistent with the idea that VLLMs memorize the training data, organizing it by similarity of association.
- GPT-3.5 is a black box. We don't know what it has been trained with (Fu *et al.*, 2022).
- It may even have been trained on our test data.

2. An Unsupervised Approach to NLI

- Build an unsupervised natural language Knowledge Graph (KG) from large amounts of multiply-authored text by extracting subject-relation-object triples by machine-reading different articles about the same events grounded in the same named-entity tuples.

- Map the KG onto a learned directed Entailment graph (EG).

- Learn from such observations that if one entity of type team *beat* another entity of that type in one document it's likely that the same two entities will *play against* each other in another.

Entailment Graphs

- We have done this in English and Chinese, using a variety of methods: (Hosseini *et al.*, 2018, 2019, 2021; Li *et al.*, 2022b).

- These methods scale: (20M sentences $\Rightarrow$ >200M sentences).

- Entailment Graphs are an efficient representation for semantics and inference using Carnap (1952) called Meaning Postulates, what Wittgenstein (1953) seems to have meant by "Meaning as Use", and what Fodor (1975) thought of as semantics.

- They can be used for inference from specific statements in a text to answers in QA.

Some Statistics on Unsupervised KG/EG

- Knowledge Graphs built on NewsSpike and NewsCrawl (Hosseini *et al.*, 2021)
  - Newspike is 0.5M multiply-sourced news articles over 2 months, 20M sentences; NewsCrawl is 5.4M articles sourced over 9 years, 256M sentences
  - NewsSpike KG has 326K typed relations, NewsCrawl, 1.05M
  - NewsSpike 29M relation triple tokens (before cutoff); NewsCrawl 729M.
  - NewsSpike 8.5M triple tokens (after cutoff); NewsCrawl 35m.
  - NewsSpike 3.9M triple types (after cutoff); NewsCrawl 13.4m
- We have built working typed global entailment graphs:
  - NewsSpike EG has 346 local typed subgraphs, NewsCrawl, 691
  - NewsSpike 23 subgraphs >1K nodes; NewsCrawl, 161
  - NewsSpike 7 subgraphs >10K nodes; NewsCrawl, 21

Statistics on Chinese KG/EG

- Chinese Knowledge Graphs built on WebHose and CLUE (Li *et al.*, 2021)
  - Webhose is 0.3M multiply-sourced news articles over 1 month, 19M sentences; CLUE is 2.4M articles sourced over 1 year, 193M sentences
  - WebHose KG has 363K typed relations, CLUE, 127M
  - WebHose 35M relation triple tokens (before cutoff); CLUE 792M.
  - WebHose 8.6M triple tokens (after cutoff); CLUE 18.5M.
  - WebHose 1.4M triple types (after cutoff); CLUE 276K
- We have built Chinese working typed global entailment graphs:
  - WebHose EG has 942 local typed subgraphs, CLUE, 384
  - WebHose 149 subgraphs $>$1K nodes; CLUE, 38
  - WebHose 26 subgraphs $>$10K nodes; CLUE, 4

Open Domain QA with Entailment Graphs

- Current work (Cheng *et al.*, 2022) uses the Newspike-based English Entailment Graph to augment a Knowledge Graph built from the entire Wikipedia corpus, and performs strongly in comparison to LMs on standard QA datasets.

Corpus	Relation		KB	LM				Our Methods	
		Freq	RE	ELMo	BERT-large	RoBERTa	Transformer-XL	KG	KG + EG
Google-RE	birth-place	4.6	13.8	-	16.1	-	2.7	19.9	**27.7**
	birth-date	1.9	1.9	-	1.4	-	1.1	7.7	**8.5**
	death-place	6.8	7.2	-	14.0	-	1.0	14.6	**26.0**
	Total	4.4	7.6	2.0	10.5	4.8	1.6	14.0	**20.7**
T-REx	Total	22.0	33.8	1.0	31.5	27.1	18.3	29.2	**35.1**
YAGO3-10	Total	-	-	1.0	2.9	1.4	1.7	5.1	**10.2**

Table 2: In zero-shot cloze-style question answering, F-score for a frequency baseline, a information extraction with entity linking (RE), ELMo, BERT-large, RoBERTa, Transformer-XL. The KG is built on the whole Wikipedia corpus and the EG means our entailment graph trained on NewsSpike.

How about the GPT3 Baselines :@?

- Used similarly, zero-shot and out-of-the-box as a Latent Entailment Graph, GPT3.5 generally scores below Unsupervised KG+EG, though better than BERT.
- However, when Retrieval-Augmented or prompted with a relevant IR snippet, GPT3.5 memorization (unsurprisingly) does terrifyingly well on these questions, which are attested in the training data.

3. Combining Entailment Graphs with LLMs

- The Problem for the directional Entailment Graph is Zipfian Sparsity of Machine-Reading.

- Can we Smooth Entailment Graphs with non-sparse but non-directional LMs without compromising the directional precision of EG?

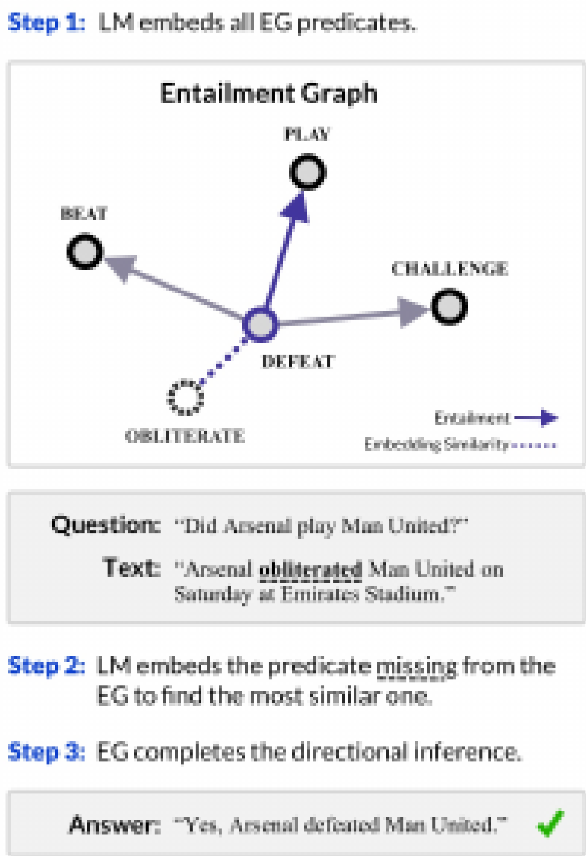
The Idea

- If the P(remise)/Antecedent and/or H(ypothesis)/Consequent are missing from the EG through sparsity, EG loses.

- If we can find P' and/or H' that are in the graph, then:
  - if $P \models P'$ and/or $H' \models H$, and
  - $P' \models H'$ is in the graph, then by transitivity of entailment:
  - $P \models H$, else:
  - $P \not\models H$.

- The idea (McKenna *et al.*, 2022): Iff P and/or H are not in the graph, use LMs to find P' and/or H' that ARE in it.

Smoothing Entailment Graphs with LMs



Smoothing Entailment Graphs with LMs

- For P and/or H that is missing in the EG find the K nearest neighbour relations P' and/or H' that are in the EG, using contextualized embedding vectors.

- Then try to establish $P/P' \models H/H'$.

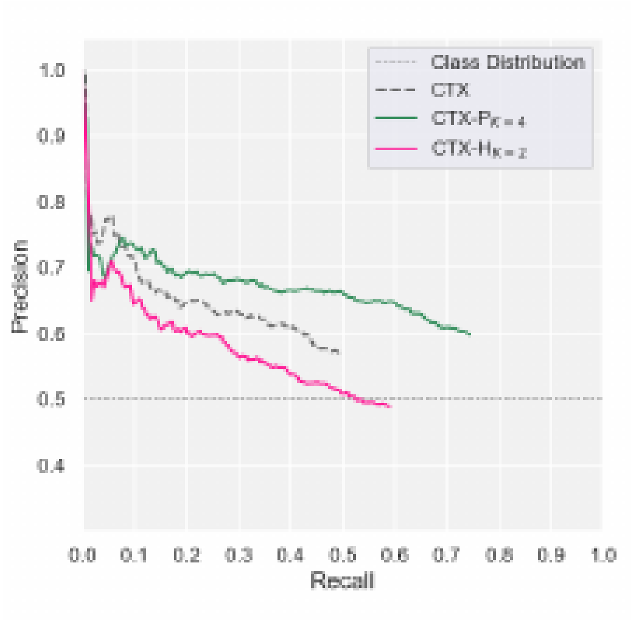
- If $P/P' \models H'/H$, assume $P \models H$

- Note that there is no guarantee for LM-KNN P' and/or H' that $P \models P'$ and/or $H' \models H$.

- Nevertheless, we are minimizing the impact on precision of the non-directional LM.

Smoothing Entailment Graphs with LMs

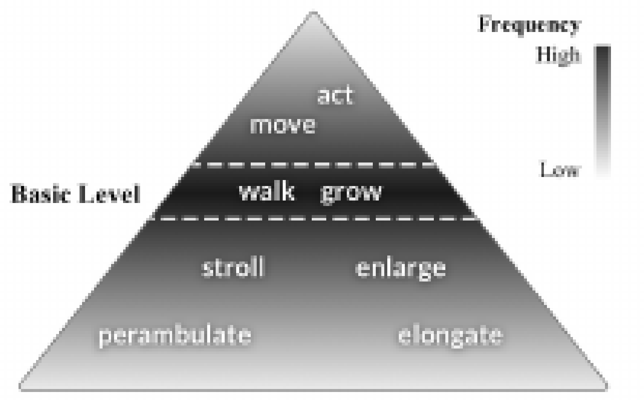
- Smoothing with an LM (RoBERTa) works for P, the antecedent:



- However, LM smoothing is deleterious for H, the consequent.
- Why is LM smoothing asymmetrical for P and H?

Why does LM Smoothing Work At All?

- There is a decrease in frequency with distance on either side of the basic level of "natural kinds" for terms on the hypernym-hyponym dimension of generality-specificity;
- There is also an increase in the number of terms with specificity:



This bias is well-known, as causing "translationese" in MT.

Why is LM Smoothing Asymmetrical?

- This skewed distribution leads to a bias towards more frequent and more general predicates in generating nearest in-graph neighbours P' and/or H' for missing P and/or H using LMs.

- Since specifics are often hyponyms and related generics hypernyms, it is likely that $P \models P'$ obtained in this way.

- However, by the same reasoning, the nearest neighbours H' of H that are most likely to be in the EG are likely to be hypernyms of H, rather than hyponyms, so that it is less likely that $H' \models H$

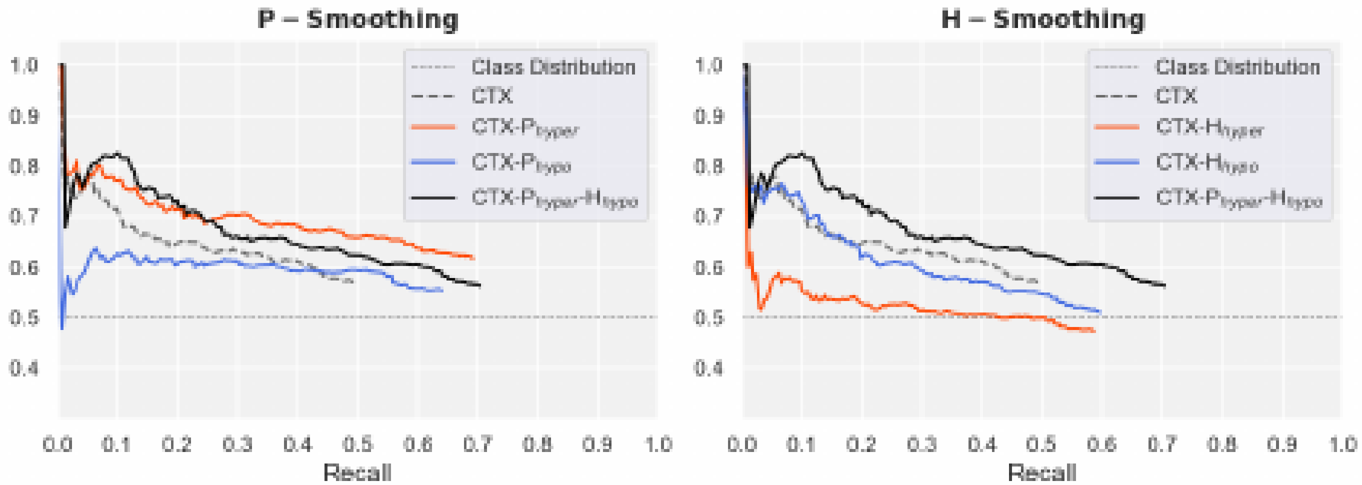
- Can we show that this is the explanation for the asymmetry?

Smoothing EGs with WordNet

- WordNet (and its mono- and multi-lingual generalization BabelNet) constitute largely distribution-neutral Hypo-Hypernym lattices.
- Use WordNet to optimally smooth P with guaranteed hypernyms and H with guaranteed hyponyms.
- McKenna *et al.* (2022) use WordNet *has hypernym* relation to identify hyper- and hypo-nyms P' and H' to smooth Hosseini *et al.* (2021) (CTX, our strongest EG).
- We test on the 2,930 question directional subset of our new ANT NLI dataset, constructed using WordNet antonyms as negative examples for comparison with supervised approaches (Bijl de Vroe *et al.*, 2022).
- The upward frequency bias still works asymmetrically for smoothing P and against smoothing H.

Smoothing EGs with WordNet

- Graphs respectively show effect of smoothing P and H with hypernyms and hyponyms against identical dashed baseline:
- They show the predicted opposite hyper/hypo effects for P and H, together with curves for predicted optimal joint P_{hyper} and H_{hypo} (identical black trace).



Smoothing EGs with WordNet

- There is the predicted hypernym facilitation in P_{hyper}.
- There is no significant hyponym facilitation for CTX in H_{hypo}.[1]
- Nevertheless, smoothing with $P_{hyper} + H_{hypo}$ significantly improves CTX over P_{hyper} alone (black trace).
- The additive effect seems to arise because, although present in EG, hyponym H' is even less frequent in text than absent H.
- It is therefore quite unlikely that EG-mining saw much evidence for $P \models H'$.
- However, P' is more frequent than P, so (given that both P' and H' are in the graph), it is a bit more likely that $P' \models H'$ is in the graph

[1] We do in fact see some H_{hypo} facilitation for our weaker EG Hosseini *et al.*, 2018.

Conclusion

- LLMs work by memorizing the training data, organized by associative similarity.

- that training data is, by definition, unlikely to include statements of entailments. (Entailments "go without saying").

- Fine-tuning LLMs on NLI datasets just seems to pick up artefacts.

- Moral: You can exploit the associative similarity of LLM neighborhoods to smooth recall in entailment graphs, without compromising their precision. . .

- . . . supporting inference needed for generation, summarization, and Open-Domain QA, as Claire always reminds us .

Thanks. . .

- To Claire for her inspiring work!
- The research was funded in part by ERC grant SEMANTAX and Huawei Edinburgh Laboratory

References

Carnap, Rudolf, 1952. "Meaning Postulates." *Philosophical Studies* 3:65–73. reprinted as Carnap, 1956:222-229.

Carnap, Rudolf (ed.), 1956. *Meaning and Necessity*. Chicago: University of Chicago Press, second edition.

Cheng, Liang, Hosseini, Javad, and Steedman, Mark, 2022. "Unsupervised Common-Sense Predicate Inference for Open-domain Question Answering." In *submitted*.

Fan, Angela, Gardent, Claire, Braud, Chloé, and Bordes, Antoine, 2019. "Using

Local Knowledge Graph Construction to Scale Seq2Seq Models to Multi-Document Inputs." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4186–4196.

Fodor, Jerry, 1975. *The Language of Thought*. Cambridge, MA: Harvard.

Fu, Yao, Peng, Hao, and Khot, Tushar, 2022. "How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources." https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc.

Geffet, Maayan and Dagan, Ido, 2005. "The Distributional Inclusion Hypothesis and Lexical Entailment." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. ACL, 107–114.

Hosseini, Javad, Chambers, Nathaniel, Reddy, Siva, Ricketts-Holt, Xavier, Cohen, Shay, Johnson, Mark, and Steedman, Mark, 2018. "Learning Typed Entailment Graphs with Global Soft Constraints." *Transactions of the Association for Computational Linguistics* 6:703–718.

Hosseini, Javad, Cohen, Shay, Johnson, Mark, and Steedman, Mark, 2019. "Duality of Link Prediction and Entailment Graph Induction." In *Proceedings of the 57th Annual Conference of the Association for Computational Linguistics (long papers)*. ACL, 4736–4746.

Hosseini, Javad, Cohen, Shay, Johnson, Mark, and Steedman, Mark, 2021. "Open-Domain Contextual Link Prediction and its Complementarity with Entailment Graphs." In *Findings of the Association for Computational Lingustics: EMNLP*. 1137–1150.

Li, Tianyi, Hosseini, Javad, Weber, Sabine, and Steedman, Mark, 2022a.

"Language Models are Poor Learners of Directional Inference." In *Findings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 903–921.

Li, Tianyi, Li, Sujian, and Steedman, Mark, 2021. "Semi-Automatic Construction of Text-to-SQL Dataset for Domain Transfer." In *Proceedings of the 14th International Conference on Parsing Technology*. 38–49.

Li, Tianyi, Weber, Sabine, Hosseini, Javad, Guillou, Liane, and Steedman, Mark, 2022b. "Cross-lingual Inference with a Chinese Entailment Graph." In *Findings of the Association for Computational Linguistics*. 1214–1233.

McKenna, Nick *et al.*, 2022. "Smoothing Entailment Graphs with Language Models." *arXiv preprint arXiv:2208.00318* .

Schmitt, Martin and Schütze, Hinrich, 2021a. "Continuous Entailment Patterns for Lexical Inference in Context." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 6952–6959.

Schmitt, Martin and Schütze, Hinrich, 2021b. "Language Models for Lexical Inference in Context." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1267–1280.

Tirumala, Kushal, Markosyan, Aram, Zettlemoyer, Luke, and Aghajanyan, Armen, 2022. "Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models." *Proceedings of the 36th Conference on Neural Information Processing Systems (NIPS)* .

Bijl de Vroe, Sander, Guillou, Liane, Johnson, Mark, and Steedman, Mark, 2022. "Temporality in General-Domain Entailment Graph Induction." In *submitted*.

Webber, Bonnie, Gardent, Claire, and Bos, Johan, 2002. "Position Statement: Inference in Question Answering." In *Proceedings of the International Conference on Language Resources and Evaluation*. Las Palmas: ELRA, 24–31.

Wittgenstein, Ludwig, 1953. *Philosophische Untersuchungen (Philosophical Investigations)*. Oxford: Basil Blackwell.

Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol, 2021. "Understanding Deep Learning (Still) Requires Rethinking Generalization." *Communications of the ACM* 64:107–115.

Zhang, Congle and Weld, Daniel, 2013. "Harvesting Parallel News Streams to Generate Paraphrases of Event Relations." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Seattle: ACL, 1776–1786.