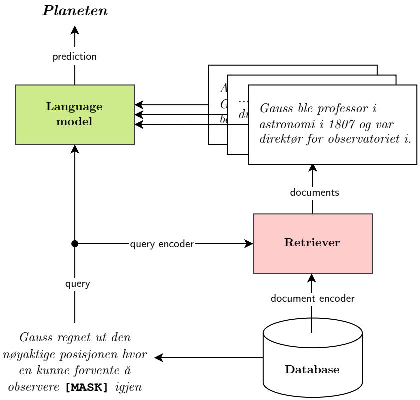
MORE ROOM FOR LANGUAGE –
INVESTIGATING THE EFFECT OF RETRIEVAL ON LANGUAGE MODELS

David Samuel, Lucas Georges Gabriel Charpentier, Sondre Wold

Language Technology Group
University of Oslo

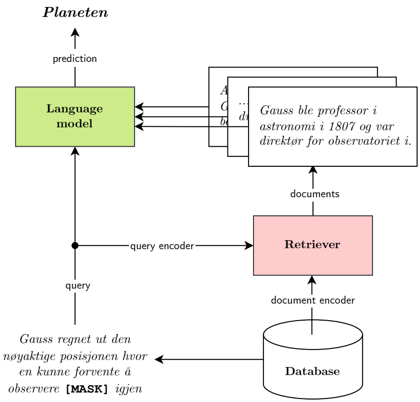
First of all – what is retrieval augmentation?

- The language model gets help during pretraining: we give it the most relevant documents.
- Very different from standard pretraining where the language model is on its own.



What is our goal?

- Our question: **How does retrieval augmentation change the behavior of a language model?**



Three dimensions of the examined 'behavior'

1. World knowledge – How many facts are stored in the weights?
2. Syntactic knowledge – More local & low-level linguistic understanding (DP)
3. Language understanding – More global & high-level understanding of language (SQuAD)

They can't be clearly separated, but we are interested in their relative change, not in the absolute values.

- For example, answering *'What is the capital of Germany?'* requires understanding the English syntax, knowledge of geography is not enough.

Fully-controllable 'ideal retrieval' methodology

- Studying realistic retrieval models comes with many variables that need to be controlled
  - What is the retriever? BM-25 or a dense model? Which dense model exactly?
  - What is the database? Wikipedia, Common Crawl or a knowledge base?
  - How do you deal with duplicates?
  - How do you chunk the documents?
  - How many documents are retrieved?
  - How are the documents fed into the language model?
  - Is the pipeline trained end-to-end or not?
- We need an ideal setting where we can fully control the retrieval accuracy and where we can abstract away from the technical details.

Fully-controllable 'ideal retrieval' methodology

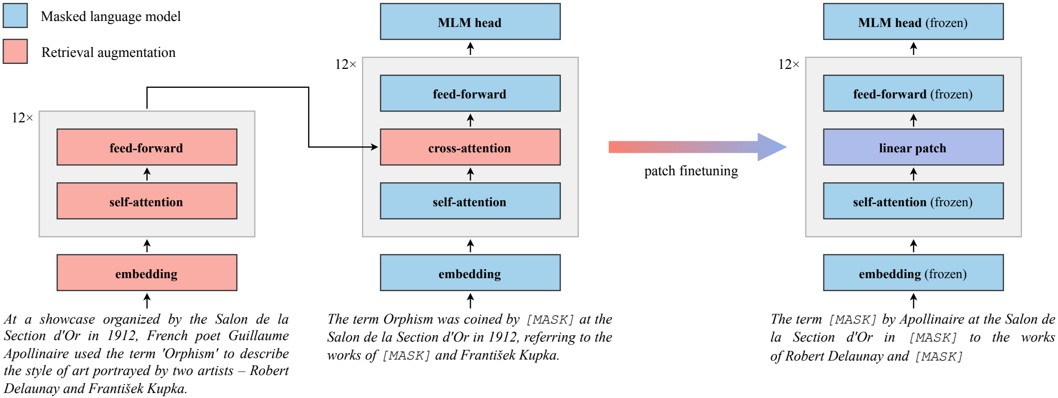
- Solution: paraphrase!
- Example:
  - Original: *"The term **Orphism** was coined by **Apollinaire** at the **Salon de la Section d'Or** in **1912**, referring to the works of **Robert Delaunay** and **František Kupka**."*
  - Paraphrase: *"At a showcase organized by the **Salon de la Section d'Or** in **1912**, French poet **Guillaume Apollinaire** used the term '**Orphism**' to describe the style of art portrayed by two artists – **Robert Delaunay** and **František Kupka**."*
- A *good* paraphraser will preserve all facts while changing the surface appearance.
  - it is a model of a perfect, 100% accurate, retriever!

Fully-controllable 'ideal retrieval' methodology

- A *good* paraphraser will preserve all facts while changing the surface appearance.
- We use `Mistral-7B-Instruct-v0.1`, is it *good*?

1. Preservation of meaning – measured by the semantic similarity of the original and paraphrase
   - The average cosine similarity is 0.88 according to `all-mpnet-base-v2`.
2. The lexical (and to some extend syntactic) similarity is evaluated by the BLEU score
   - The average BLEU score is 0.13 for the raw pairs, and 0.07 for pairs with removed named entities and digits

Evaluating a stand-alone language model



Pretraining

- All models are trained from scratch on the top 10% most visited English Wikipedia pages.
  - We do this so that our dataset is rich in world knowledge.
  - We use `Mistral-7B-Instruct-v0.1` to paraphrase the Wikipedia passages.
- We test only masked language models
  - Easier to evaluate and work nicely at 'small' scale
- Three sizes – `base` (98M), `small` (28M), and `x-small` (9M)
- The `base` model is also pretrain with 25% and 50% noisy retrieval
  - To simulate a more realistic scenario

Evaluation – world knowledge: LAMA probes

- A probe to test the factual and commonsense knowledge in language models, introduced in Petroni et al. (2019).
- The test uses cloze-style statements as an evaluation framework. E.g:
  - A joke would make you want to ___
  - The official language of Mauritius is ___
- Results are reported as the average precision at *k* for different values of *k*, together with the Mean Reciprocal Rank. For a given fact, we count it as correctly predicted if the object is ranked among the top *k* results, false otherwise.
- Three subsets:
  - ConceptNet
  - TREx
  - SQuAD

Evaluation – syntactic knowledge: linear probing

- Tests how much information about dependencies can be extracted from the hidden representations with a simple linear transformation.
    - A model with a better syntactic understanding should encode more of the syntactic information in the latent vectors.
    - And this information should be easily accessible (linearly separable) to be used in self-attention.
- Freeze a language model $\rightarrow$ do dependency parsing without nonlinearities $\rightarrow$ measure LAS

Evaluation – syntactic knowledge: attention probing

- We mostly follow Raganato and Tiedemann (2018), and Ravishankar et al. (2021) in their evaluation setup of attention probing.
- The goal is to decode dependency trees directly from the attention weights.
  - With the idea that a language model with better syntactic understanding should better utilize the hierarchical syntactic structure in its attention mechanism.
- Take a matrix with attention probabilities $\rightarrow$ make it symmetric $\rightarrow$ find the maximum spanning tree $\rightarrow$ measure UUAS

Evaluation – syntactic knowledge: BLiMP

- Zero-shot linguistic acceptability judgments by Warstadt et al. (2020a).
- Consists of 67 tasks, each focuses on a specific linguistic feature, which is tested with 1 000 sentence pairs.
- Each pair of sentences differs minimally on the surface level, but only one of the sentences is grammatically valid.
- We test if the LM assigns a higher (pseudo-)probability to the correct sentence
- a) The cats annoy Tim. *(grammatical)*
b) The cats annoys Tim. *(ungrammatical)*

Evaluation – syntactic knowledge: MSGS

- A finetuning task that aims to ascertain whether a model biases surface or linguistic features by Warstadt et al. (2020b)
- Finetuned on ambiguous data, containing both feature types or neither while evaluation is done on unambiguous data with labels indicating the presence of the linguistic feature.
- Adapting the Mathews' Correlation Coefficent scoring such that a score of -1 = Surface bias, 1 = Linguistic bias.
- Surface features include lexical content, relative token position, absolute token position, orthography, and length.
- Linguistic features include main verb form, syntactic category, control raising, and morphology.

- a) The cat chased a mouse. *Relative token position: positive*
  b) A cat chased the mouse. *Relative token position: negative*

Evaluation – language understanding: LAMBADA

- A zero-shot language modeling tasks that focuses on resolving long-range dependencies in text (Paperno et al., 2016).
- While it has been traditionally used for evaluating autoregressive LMs, we adapt the task for masked language models.
- *Preston had been the last person to wear those chains, and I knew what I'd see and feel if they were slipped onto my skin – the Reaper's unending hatred of me. I'd felt enough of that emotion already in the amphitheater. I didn't want to feel anymore. "Don't put those on me," I whispered. "Please." Sergei looked at me, surprised by my low, raspy please, but he put down the* ***{answer}****.*

Evaluation – language understanding: LAMBADA

- A zero-shot language modeling tasks that focuses on resolving long-range dependencies in text (Paperno et al., 2016).
- While it has been traditionally used for evaluating autoregressive LMs, we adapt the task for masked language models.
- *Preston had been the last person to wear those chains, and I knew what I'd see and feel if they were slipped onto my skin – the Reaper's unending hatred of me. I'd felt enough of that emotion already in the amphitheater. I didn't want to feel anymore. "Don't put those on me," I whispered. "Please." Sergei looked at me, surprised by my low, raspy please, but he put down the* ***{answer}****.*
  Gold answer: *chains*

Evaluation – language understanding: GLUE

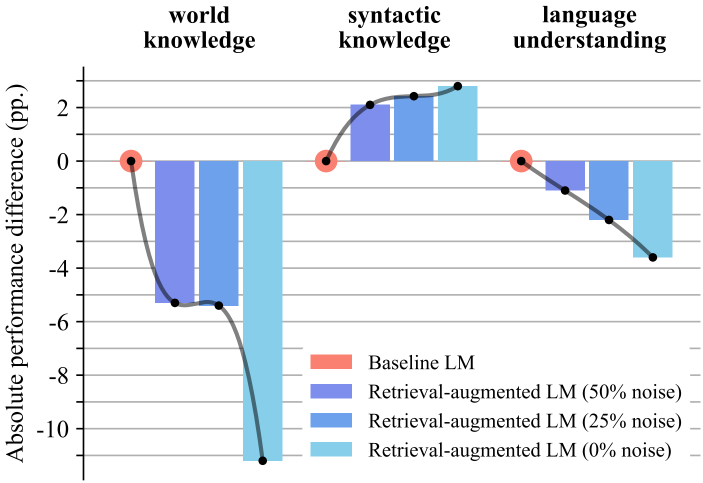
- A finetuning benchmark that evaluates multiple downstream tasks, put together by Wang et al. (2019).
- The benchmark evaluates language acceptability, paraphrase recognition, natural language inference, and sentiment analysis.
- 4 different metrics are used to evaluate the tasks (one per task): Mathews' Correlation Co-efficent, F1-score, Accuracy, and Pearsons' / Spearman's r-score.

Evaluation – language understanding: SQuAD

- Lastly, we use the Stanford Question Answering Dataset (SQuAD), a reading comprehension task (Rajpurkar et al., 2016).
- Models are given two inputs: a question, and a longer passage. The task is to predict the span of the passage that answers the question.
- 100 000+ questions, created from Wikipedia using crowd-sourcing.

Answer type	Percentage
Common noun phrase	31.8%
Other entity	15.3%
Person	12.9%
Other numeric	10.9%
Date	8.9%
Verb phrase	5.5%
Location	4.4%
Adjective phrase	3.9%
Clause	3.7%
Other	2.7%

Results – high-level view



Results – medium-level view

Model	world knowledge			syntactic knowledge				language understanding		
	ConceptNet (MRR ↑)	SQuAD (MRR ↑)	TREx (MRR ↑)	linear probing (LAS ↑)	attention probing (UUAS ↑)	BLiMP (Acc. ↑)	MSGS (LBS ↑)	LAMBADA (Acc. ↑)	GLUE (Avg. ↑)	SQuAD (F_1 ↑)
reference model (110M)										
bert-base-cased	*26.0*	*34.0*	*62.0*	*82.0*	*45.1*	*85.6*	*-0.10*	*44.8*	*82.1*	*88.4*
base (98M)										
− retrieval	**20.3**	**32.1**	**53.6**	78.1	48.0	82.9	-0.47	**46.0**	**82.2**	**91.2**
+ retrieval (50% noise)	17.7	23.2	49.1	79.8	51.3	81.3	**-0.37**	43.2	82.0	90.7
+ retrieval (25% noise)	18.1	23.4	48.3	79.9	51.6	82.7	-0.38	40.6	81.9	90.2
+ retrieval (0% noise)	14.9	15.8	41.5	**80.2**	**51.8**	**83.2**	**-0.37**	37.5	81.2	89.7
small (28M)										
− retrieval	**17.2**	**28.3**	**47.4**	71.1	49.7	78.6	-0.56	**35.1**	78.0	**88.6**
+ retrieval	11.8	15.3	36.3	**71.2**	**50.4**	**78.8**	**-0.53**	26.2	**78.4**	86.2
x-small (9M)										
− retrieval	**9.9**	**14.7**	**39.2**	63.3	45.5	**73.4**	**-0.55**	**25.3**	75.2	**81.1**
+ retrieval	7.5	10.6	23.4	**63.6**	**49.2**	73.3	-0.57	19.3	**76.0**	78.7

Discussion: Retrieval augmentation separates linguistic knowledge from world knowledge

- There is clear trend between the world knowledge tasks and linguistic tasks:
  - When a LM can rely more on retrieval, it remembers less facts and gets progressively worse on all evaluated world knowledge tasks.
  - On the other hand, its syntactic understanding gets consistently better.
- LM with retrieval does not allocate as many parameters to store world knowledge and instead uses them for other features, such as syntax.
- Thus, retrieval-augmented pretraining leads to separation between the world knowledge (in the retriever) and syntactic knowledge (in the language model).
- Retrieval-based pretraining can be a promising avenue for efficient language modeling.

Discussion: Retrieval augmentation negatively impacts NLU performance

- Contrary to syntactic understanding, the language understanding gets worse with retrieval-augmented pretraining.
- The fine-grained GLUE results show that this affects tasks that require global inter-sentence comprehension tasks (NLI) more than the short-range local tasks (CoLA or SST-2).
- We argue that this is in part caused by the lacking factual knowledge but it is also indirectly caused by the mechanism of retrieval-augmented pretraining.
  - When looking for the global context, the language model is incentivized to trust the retrieved document more than the partially masked input.
- This poses a challenge to utilizing retrieval-augmentation for pretraining general-purpose language models.

Discussion: Poor retrieval quality does not negatively impact pretraining

- Noisy retrieval pretraining does not lead to an overall drop in performance.
- Instead, it interpolates the behavior of standard pretraining and of pretraining with a perfect retrieval.
- Our results suggest that a subpar (but computationally inexpensive) retrieval should not negatively impact training.

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, N. Q., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Raganato, A. and Tiedemann, J. (2018). An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ravishankar, V., Kulmizev, A., Abdou, M., Søgaard, A., and Nivre, J. (2021). Attention can reflect syntactic structure (if you let it). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3031–3045, Online. Association for Computational Linguistics.

References ii

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of ICLR*.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020a). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Warstadt, A., Zhang, Y., Li, X., Liu, H., and Bowman, S. R. (2020b). Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.