BERT-kNN: Adding a kNN Search Component to Pretrained Language Models for Better QA

Nora Kassner, Hinrich Schütze
Center for Information and Language Processing (CIS)
LMU Munich, Germany
kassner@cis.lmu.de

Abstract

Khandelwal et al. (2020) show that a k-nearest-neighbor (kNN) component improves language modeling performance. We use this idea for open domain question answering (QA). To improve the recall of facts stated in the training text, we combine BERT (Devlin et al., 2019) with a kNN search over a large corpus. Our contributions are as follows. i) We outperform BERT on cloze-style QA by large margins without any further training. ii) We show that BERT often identifies the correct response category (e.g., central European city), but only kNN recovers the factually correct answer (e.g., "Vienna").

1 Introduction

Pre-trained language models (PLMs) like BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019) have emerged as universal tools that not only capture a diverse range of linguistic but also (as recent evidence seems to suggest) factual knowledge.

Petroni et al. (2019) introduced LAMA (LAnguage Model Analysis) to investigate PLMs' capacity to recall factual knowledge without the use of fine-tuning. Since the PLM training objective is to predict masked tokens, question answering tasks can be reformulated as cloze questions; e.g., "Who wrote 'Ulysses'?" is reformulated as "[MASK] wrote 'Ulysses'." In this setup, Petroni et al. (2019) show that, on QA, PLMs outperform baselines trained on automatically extracted knowledge bases.

Still, given that PLMs have seen more data than any human could read in a lifetime, their performance on open domain QA seems poor. Even LAMA facts that PLMs do get right are not necessarily "recalled" from the training experience as many of them are easy-to-guess (Poerner et al., 2019). Choosing BERT as our PLM, we therefore introduce BERT-kNN in this paper (see Figure 1): BERT-kNN combines BERT's predictions with a kNN search over a text collection where the text collection can be BERT's training set or any other suitable text corpus. Due to its kNN component and its resulting ability to directly access facts stated in the searched text, BERT-kNN outperforms BERT on cloze-style QA by large margins.

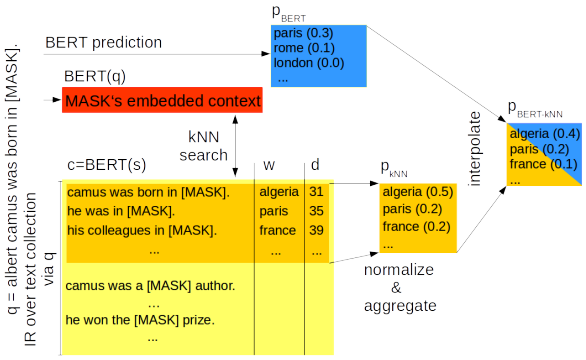

Figure 1: Schematic depiction of BERT-kNN: BERT's prediction for query q are interpolated with a kNN-search component. The query q is input to an IR step. The BERT embeddings of the retrieved contexts $BERT(s)$ together with the target word build a key-value datastore $c - w$ (yellow). The kNN search runs between the BERT embeddings of the query BERT(q) (red) and the c of the datastore. The corresponding w of the kNNs and their distances d are returned (orange). The are aggregated and normalized. Finally, the predictions of the kNN-search component and BERT's predictions are interpolated.

In more detail, we use BERT to embed each token's context in the text collection. Each pair of context embedding and token is stored as a key-value pair in a datastore. At test time for a cloze question q, the MASK's embedded context serves as query $BERT(q)$ to find the k context-target pairs in the datastore that are closest. To make this more effective, we first query a separate information retrieval (IR) index with the original question

q and only search over the top m hits when finding the k nearest neighbors of $BERT(q)$ in embedding space. The final prediction is an interpolation of the kNN search and the PLM predictions.

We find that the PLM often correctly predicts the answer category and therefore the correct answer is often among the top k nearest neighbors. A typical example is "Albert Einstein was born in [MASK]": the PLM knows that a city is likely to follow and maybe even that it is a German city, but it fails to pick the correct city. On the other hand, the top-ranked answer in the kNN search is "Ulm" and so the correct filler for the mask can be identified.

BERT-kNN outperforms BERT on the LAMA cloze style QA dataset without any further training. Even though BERT-kNN is based on BERT-base, it also outperforms BERT-large on 3 out of 4 LAMA subsets. The performance gap between BERT and BERT-kNN is most pronounced on hard-to-guess facts. As this method can be applied to any kind of text collection (not just the PLM training corpus), BERT-kNN can potentially correctly give answers that BERT has never seen in its training corpus.

2 Data

The LAMA dataset is a cloze style QA dataset that allows to query PLMs for knowledge base like facts. A cloze question is generated from a subject-relation-object triple from a knowledge base and from a templatic statement for the relation that contains variables X and Y for subject and object (e.g, "X was born in Y"). The subject is substituted for X and [MASK] for Y. The triples are chosen such that Y is always a single-token answer.

LAMA covers different sources: The GoogleRE[1] set covers the three relations "place of birth", "date of birth" and "place of death". T-REx (ElSahar et al., 2018) consists of a subset of Wikidata triples covering 41 relations. ConceptNet (Li et al., 2016) combines 16 commonsense relationships between words and phrases. The underlying Open Mind Common Sense corpus provides matching statements to query the language model. SQuAD (Rajpurkar et al., 2016) is a standard question answering dataset. LAMA contains a subset of 305 context-insensitive questions and provides manually reformulated cloze-style questions to query the model.

Poerner et al. (2019) introduce LAMA-UHN, a subset of LAMA's T-REx and GoogleRE questions from which easy-to-guess facts have been removed.

Corpus	BERT-base	BERT-large	BERT-kNN
LAMA	27.7	30.6	**36.8**
LAMA-UHN	20.6	23.0	**31.8**

Table 1: Mean precision at one (P@1) for LAMA and LAMA-UHN on the TREx and GooglRE subsets.

3 Method

BERT-kNN combines Bert-base with a kNN search component. We now describe the architecture of BERT-kNN.

BERT. This method is applicable to any kind of PLM. We use BERT-base-uncased (Devlin et al., 2019) as our PLM since it is top performer on LAMA. BERT estimates the probability of a masked word given it's context. BERT is pre-trained on the BookCorpus (Zhu et al., 2015) as well as a crawl of English Wikipedia. During pre-training, BERT randomly masks positions and learns to fill the words.

Datastore. Our text collection C is the 2016-12-21 English Wikipedia.[3] For each single-token word occurrence w in a sentence s in C, we compute the pair (c, w) where c is a context representation of s computed by BERT. We find that masking the occurrence of w in s and using the embedding of the masked token is an effective context representation c. We store all pairs (c, w) in a key-value datastore D where c serves as key and w as value.

Information Retrieval. We found that just using the datastore D does not give good results. We therefore use (Chen et al., 2017)'s IR system to first select a small subset of D using a keyword search. The IR index contains all Wikipedia articles. An article is represented as a bag of words and word bigrams. If the subject in question is specified we use it as-is to query the IR index, otherwise, the cloze-style question q (the [MASK] token is removed) is used. Finally, we find the top 5 relevant Wikipedia articles using TF-IDF search.

Inference. At test time, we first run the IR search between the cloze question q and datastore D and then only consider the subset of D that corresponds to the top 5 relevant Wikipedia articles. For the kNN search q is embedded in the same way as the context representations c in D: we set $BERT(q)$ to the embedding computed by BERT

[1] https://code.google.com/archive/p/relation-extraction-corpus/

[3] https://dumps.wikimedia.org/enwiki/latest/

Corpus	Relation	Statistics		model			
		Facts	Rel	BERT-base[2]	BERT-large[2]	kNN	BERT-kNN
Google-RE	birth-place	2937	1	14.9	16.1	45.5	**46.4**
	birth-date	1825	1	1.5	1.4	39.5	**39.7**
	death-place	765	1	13.1	14.0	**39.1**	38.7
T-REx	1-1	937	2	68.0	74.5	72.7	**78.0**
	N-1	20006	23	32.4	34.2	29.8	**37.4**
	N-M	13096	16	24.7	24.3	234.9	**30.1**
ConceptNet	Total	11458	16	15.6	**19.2**	4.7	13.8
SQuAD	Total	305	-	14.1	17.4	**25.9**	25.6

Table 2: Mean precision at one (P@1) for BERT-base, BERT-large, the k-NN search and the interpolation between BERT and the k-NN search (BERT-kNN) across the set of evaluation corpora.

for [MASK]. We then retrieve the k nearest neighbors of $BERT(q)$ in the 5-Wikipedia-article subset of D where $k = 512$. We convert the distances between $BERT(q)$ and the 512 nearest neighbors to a probability distribution using softmax normalization. Since a word w can occur several times in the 512 nearest neighbors, we compute its final output probability as the sum over all occurrences. Not occurring words have zero probability.

In the final step the probability distributions of BERT and the kNN search are interpolated with interpolation parameter λ (set to 0.6).

4 Evaluation

As Petroni et al. (2019) we report mean precision at rank k (P@k). P@k is 0 or 1 depending on if the true answer occurs among the the top k predictions. Averaging is done first within each relation and then across relations.

5 Results and Discussion

BERT-kNN outperforms BERT on the LAMA dataset. It obtains over 10 precision points gain over BERT-base and large. Note that our model uses BERT-base only. Table 1 shows that the performance gap between original BERT and BERT-kNN becomes even larger when evaluating on LAMA-UHU, a subset of LAMA with hard to guess facts.

Table 2 shows performance on different LAMA subsets. We see that BERT-kNN outperforms BERT-base and BERT-large on 3 out of 4 LAMA subsets. On ConceptNet it shows competitive results. Huge gains are obtained on the GoogleRE dataset. Figure 2 shows precision at 1, 5 and 10. BERT-kNN performs better in all three categories.

Table 2 also shows that neither BERT nor the kNN search alone are sufficient for good performance. Only the interpolation of the two yields optimal results. In many cases, the knowledge recalled by BERT and the kNN is complementary. BERT is much better on ConceptNet relations. This seems to be due to the limitations of knowledge expressed in Wikipedia articles. Note that the interpolation parameter is kept constant over all datasets. The prediction probabilities are well calibrated in a sense that BERT-kNN is able to distinguish when to rely more on BERT or the kNN predictions.

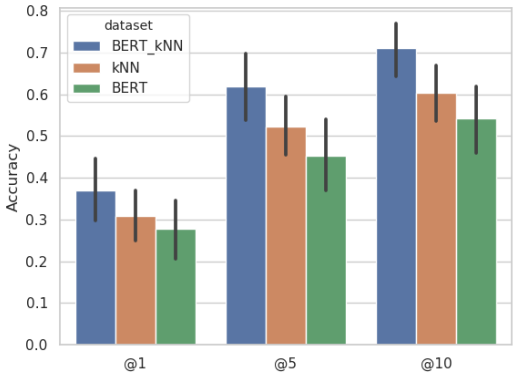

Figure 2: Mean Precison@1, Precison@5, Precison@10 on LAMA for original BERT and BERT-kNN

Table 3 compares exemplary differences in BERT and BERT-kNN predictions. We see that original BERT is good in predicting the answer category required for completing the cloze query but only the kNN-search is able to recover the actual fact.

6 Related work

PLMs are top performers for many tasks, including QA (Kwiatkowski et al., 2019; Alberti et al., 2019). Petroni et al. (2019) introduced the LAMA cloze-style QA task to query PLM's performance on knowledge base like facts. Bosselut et al. (2019)

[3]Note that the results for BERT-base and BERT-large are taken from (Petroni et al., 2019) where a slightly smaller subset of Bert's original vocabulary is used.

	Query and True Answer	Generation
Google RE	Hans Gefors was born in [MASK]. True: Stockholm	BERT-kNN: Stockholm (0.62), Oslo (0.08), Copenhagen (0.7) BERT: Oslo (0.22), Copenhagen (0.18), Bergen (0.09) kNN: Stockholm (0.97), Lund (0.02), Hans (0.0)
	Aglaja Orgeni died in [MASK]. True: Vienna	BERT-kNN: Vienna (0.61), Bucharest (0.08), Paris (0.03) BERT: Bucharest (0.19), Paris (0.08), Budapest (0.04) kNN: Vienna (1.0), 1886 (0.0), Munich (0.0)
TREx	Regiomontanus works in the field of [MASK]. True: Mathematics	BERT-kNN: Mathematics (0.25), Astronomy (0.17), Medicine (0.04) BERT: Medicine (0.09), Law (0.05), Physics (0.03) kNN: Mathematics (0.40), Astronomy (0.28), Literature (0.04)
	The headquarter of interpol is in [MASK] . True: Lyon	BERT-kNN: Lyon (0.52), Paris (0.05), Singapore (0.04) BERT: Paris (0.12), London (0.08), Brussels (0.05) kNN: Lyon (0.86), Singapore (0.07), Oslo (0.01)
ConceptNet	Ears can [MASK] sound. True: hear	BERT-kNN: hear (0.22), detect (0.16), produce (0.11) BERT: hear (0.28), detect (0.06), produce (0.04) kNN: detect (0.23), hear (0.19), produce (0.15)
	Regret is an [MASK]. True: emotion	BERT-kNN: emotion (0.1), action (0.03), evolutionary (0.02) BERT: emotion (0.25), option (0.04), art (0.04) kNN: action (0.04), evolutionary (0.03), explanation (0.03)
Squad	[MASK] is needed to pack electrons densely together. True: energy	BERT-kNN: it (0.20), energy (0.05), this (0.04) BERT: it (0.5), this (0.1), energy (0.07) kNN: energy (0.04), electrons (0,02), material (0.01)
	The capital of the ottoman empire was [MASK]. True: Istanbul	BERT-kNN: Istanbul (0.32), Constantinople (0.25), Vienna (0.02) BERT: Constantinople (0.48), Istanbul (0.33), Acre (0.02) kNN: Istanbul (0.3), Constantinople (0.1), Vienna (0.02)

Table 3: Examples of generation for BERT-base, kNN, BERT-kNN. The last column reports the top three tokens generated together with the associated probability (in brackets).

investigate PLMs' common sense knowledge only.

DRQA (Chen et al., 2017) is a popular open-domain QA model that combines an IR step with a neural reading comprehension model. Even tough we use the same IR module our model differs significantly. DRQA does not predict masked tokens but extracts answers from text. It does not do use PLM Transformers nor a kNN search module. But most notably BERT-kNN is fully unsupervised and does not require any extra training.

Extended work on knowledge in PLM focuses on injecting knowledge into BERT's encoder. ERNIE (Zhang et al., 2019) and KnowBert (Peters et al., 2019) are entity-enhanced versions of BERT. They introduce additional encoder layers that are integrated into BERT's original encoder by expensive further pre-training. Our approach on the other hand is not limited to labeled entities nor does it require any further training. (Poerner et al., 2019) injects factual entity knowledge into BERT's embeddings without further training but by aligning Wikipedia2Vec entity vectors (Yamada et al., 2016) with BERT's word piece vocabulary. This approach is also limited to labeled entities. Our approach is conceptually very different from entity-enhanced versions of BERT and could potentially be combined with any of the mentioned ones.

BERT-kNN architecture is based on (Khandelwal et al., 2020) where an interpolation of a PLM and a kNN search is used for language modeling. In contrast this work analyses QA. Architecturally we introduce an IR step into the model that is essential for factual correctness. We also change the hidden state used for the kNN to the masked token embeddings.

Other work that store previous hidden states in memory are Grave et al. (2016); Merity et al. (2017). They only consider recent history making it easier to copy rare vocabulary items from the recent past. They do not use PLM Transformer architecture. Again these models evaluate on LM and not on factual correctness.

7 Conclusion

This work introduced BERT-kNN, an interpolation of BERT predictions with a kNN search for unsupervised cloze style QA. BERT-kNN sets new state of the art on the LAMA dataset with top performance on hard to guess without any further training. This method potentially allows querying LMs for knowledge outside of the training domain with no additional training.

References

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT baseline for the natural questions. *ArXiv*, abs/1901.08634.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hady ElSahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*.

Edouard Grave, Armand Joulin, and Nicolas Usunier. 2016. Improving neural language models with a continuous cache. *ICLR*, abs/1612.04426.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations (ICLR)*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455, Berlin, Germany. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 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.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. Bert is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised qa. *ArXiv*, abs/1911.03681.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259, Berlin, Germany. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies

and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.