

KGI: An Integrated Framework for Knowledge Intensive Language Tasks

Md Faisal Mahbub Chowdhury*, Michael Glass*, Gaetano Rossiello,
Alfio Gliozzo and Nandana Mihindukulasooriya

{mchowdh, mrglass, gliozzo}@us.ibm.com, {gaetano.rossiello, nandana}@ibm.com
IBM Research AI, Yorktown Heights, NY, USA

Abstract

In this paper, we present a system to showcase the capabilities of the latest state-of-the-art retrieval augmented generation models trained on knowledge-intensive language tasks, such as slot filling, open domain question answering, dialogue, and fact-checking. Moreover, given a user query, we show how the output from these different models can be combined to cross-examine the outputs of each other. Particularly, we show how accuracy in dialogue can be improved using the question answering model. We are also releasing all models used in the demo as a contribution of this paper. A short video demonstrating the system is available at <https://ibm.box.com/v/emnlp2022-demo>.

1 Introduction

Recently, we proposed Re²G (Glass et al., 2022), the core of our KGI (Knowledge Graph Induction) system. Re²G combines both neural initial retrieval and reranking into a BART-based sequence-to-sequence generation. We show that the end-to-end reranking component also permits merging retrieval results from sources with incomparable scores, enabling an ensemble of BM25 and neural initial retrieval. Moreover, to train our system end-to-end, we introduce a novel variation of knowledge distillation to train the initial retrieval, reranker, and generation using only ground truth on the target sequence output. We find large gains in four diverse tasks: zero-shot slot filling, question answering, fact-checking, and dialog, with relative gains of 9% to 34% over the previous state-of-the-art on the KILT leaderboard (Petroni et al., 2021)¹.

In this work, we describe the complete KGI system, which is an enhancement of our previous work. We demonstrate how users can asynchronously interact with the system in real-time, not only for

*These authors contributed equally to this work.

¹<https://eval.ai/web/challenges/challenge-page/689/overview>

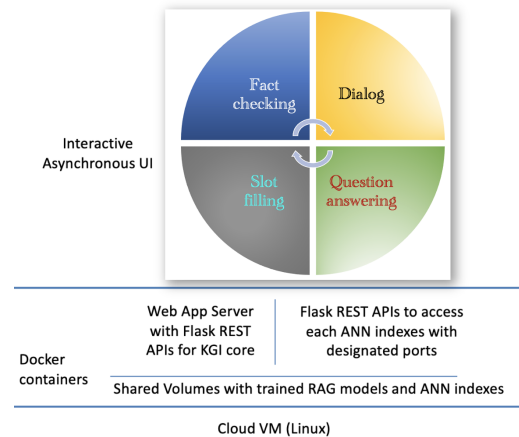


Figure 1: KGI: System Architecture

completing triples (aka slot filling), but also for dialogue, fact-checking, and open-domain question answering. We empirically show that our system is the state of the art for these tasks on the KILT leaderboard. In addition, we show how dialog accuracy can be improved by exploiting the question answering model, a novel approach demonstrated in this paper.

There are several different intended usages of our system. For example, KGI allows users to interact with different levels of verbosity. Also, it enables users to cross-examine results through different KILT tasks that are part of the same GUI.

We are releasing our best KGI core models (i.e. Re²G) that we used in this paper at <https://huggingface.co/ibm>.

2 System Architecture

The KGI system (Figure 1) is a web-based application that enables users to asynchronously interact with the system in real-time and allows users to obtain results from four different task-specific models simultaneously in different tabs in the GUI. These models are trained using Re²G model as shown in Figure 2. There is a corresponding ANN (Approxi-

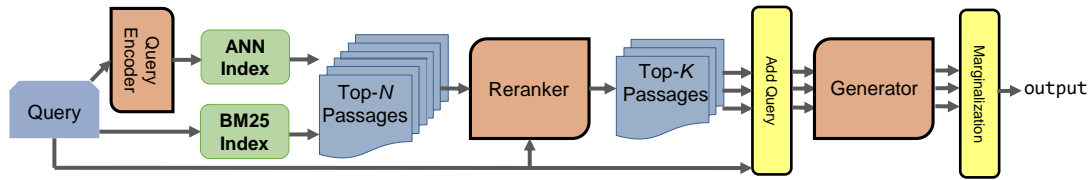


Figure 2: Re2G (Retrieve, Rerank, Generate) model (Glass et al., 2022)

mate Nearest Neighbors) index, in this case HNSW (Hierarchical Navigable Small World) (Malkov and Yashunin, 2018) using the open source FAISS library (Johnson et al., 2017)². These indexes contain the passage vectors for the source corpus of the corresponding tasks. More details about the KGI core models, particularly how a model can be trained and a corresponding index can be created, can be found in Glass et al. (2021, 2022).

The KGI core model, such as Re²G, takes a textual query as input and returns a set of generated texts as results together with a set of passages as supporting evidences/references for each of the tasks such as slot filling or fact-checking. Refer to Table 1 for examples.

2.1 Dialog supported by QA

There are two settings for dialog from the GUI. “conventional-dialog” is solely based on the KGI dialog model, while in the “hybrid” settings the system also interacts with the KGI QA model depending on the comments entered by the user. The system uses a simple Convolutional Neural Networks (CNN) based text classification model to detect whether the latest comment entered by the user is a question. If a comment is identified as a question, and if it contains at least one noun phrase without a pronoun or adverb, the system creates a query by appending all such noun phrases from the previous user utterances in the current dialog history (with full-stop as separators) with the question and pass it to the QA model. If none of the tokens in the best-ranked answer provided by the QA model is part of the same dialog history, the system picks the QA answer (and corresponding evidences) as the response for the dialog.

3 Application to Diverse NLP Tasks

3.1 The tasks

As mentioned earlier, we demonstrate the robustness of our system on four NLP tasks that are

²<https://github.com/facebookresearch/faiss>

part of the KILT leaderboard. Among them, fact-checking requires deep knowledge about the claim and reasoning over multiple documents. In slot filling, the goal is to collect information on certain relations of entities. For the open domain QA, the goal is producing the correct answer for a question after reasoning over an entire knowledge source (in this case, Wikipedia), without a predefined location for the answer. Finally, for dialog the goal of the system is to engage in chitchat, relying on topical and factual knowledge, on a wide array of (non-specified) topics with a user. There is another task, entity linking, in KILT in which we did not participate.

The KILT benchmark consists of eleven datasets spanning five distinct tasks. All these task-specific datasets in this benchmark are grounded in the same snapshot of Wikipedia. We refer the readers to Petroni et al. (2021) for details about the datasets. Table 1 shows the input and output types for the four different tasks considered.

3.2 Application of KGI

The KGI system, although originally designed for zero-shot slot filling, is based on a very general approach: conditional generation with retrieval. An input text is used to retrieve passages from a corpus of knowledge, then a generation component conditions both the input text and the returned passages to produce an output text.

The KILT benchmark was introduced to evaluate the capabilities of pre-trained language models to address NLP tasks that require access to external knowledge. As mentioned by the organizers, developing general models for such knowledge-intensive tasks is difficult as each task might require computationally expensive indexing of custom knowledge sources, in addition to dedicated infrastructure. So, it is a perfect playground to verify the generalizability and robustness of KGI.

Training models for each of the above tasks are carried out in two phases: DPR training and generation training. The training procedure and hyperpa-

rameters are exactly the same as described in our earlier works (Glass et al., 2021, 2022)³⁴.

The slot filling dataset, T-REx (Elsahar et al., 2018), provides as input a head entity and relation, and expects as output the entity or term that fills the slot, also called the tail entity. The T-REx dataset contains 2.3M instances. We use only 370k training instances by down-sampling the relations that occur more than 5000 times. This reduces the training time required while keeping state-of-the-art performance. The development and test sets each have 5k instances.

The question answering datasets are “open” versions of Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). Unlike the original versions, the relevant Wikipedia page must be found by a retrieval step. The training sets for Natural Questions and TriviaQA contain 87k and 62k questions, with another 3k and 5k for the development and 1.4k and 6.5k for test.

The fact-checking dataset in KILT is FEVER (Fact Extraction and VERification). It is a combination of the two FEVER versions (Thorne et al., 2018, 2019) omitting the NOTENOUGHINFO class. There are approximately 10k instances in the development and test sets, and 100k for training. FEVER is a classification task, but we cast it as a generation task by training the model to generate either the token “SUPPORTS” or “REFUTES”.

Wizard of Wikipedia (Dinan et al., 2018) is the dialog dataset. The input is a short dialog history ending with the information seeker’s turn. The expected output is a fact presented conversationally or just an utterance or question mentioning content from a relevant Wikipedia page. It is the smallest dataset with approximately 3k instances in development and test and 64k in train.

3.3 Results

Table 2 shows the results of our system on KILT datasets for different tasks. At the time of the submission in 2021, our earlier version of KGI core models (namely, KGI₀ and KGI₁) achieved the best results in the KILT leaderboard. Our new KGI core models, Re²G, achieves significantly better results. In fact, it considerably advanced the state-of-the-art across five KILT datasets, and still holds the top position in four of the five. Particularly, our system architecture permits us to ensemble DPR and

³<https://github.com/IBM/kgi-slot-filling>

⁴<https://github.com/IBM/kgi-slot-filling/tree/re2g>

BM25, which is enabled by our incorporation of a reranker, further improving accuracy. Our online knowledge distillation is able to improve the performance of DPR in four of the five datasets, despite the loss in end-to-end training not depending on the DPR scores.

4 Examples and Analysis

4.1 Complementing information from different applications

As mentioned earlier, one of our goals is to allow a user to interact with different levels of verbosity and then cross-examine the results to check whether response from one application (e.g. fact checking) supports response from another application (e.g. dialog). This can be checked not only by looking at the responses but also through the accompanying evidences. Figure 3 shows such an example where the user intends to know the host of the 2014 Soccer World Cup. This has been formulated in different ways according to the application. All four KGI models for the corresponding applications provided the correct answer (Brazil). Note, all the models use the Wikipedia corpus (as provided by the KILT organizers) as the knowledge source, yet the corresponding supporting evidences are not always the same.

In real world scenarios, the ability to cross-check information and compare complementing evidence is important for decision making, specially for subject matter experts.

4.2 Dialog by exploiting results of open domain QA

Our view is that the most natural choice to automatically combine results from different task specific models and improve results of a particular task is dialog. So, we created a hybrid settings for the dialog application as described in Section 2.1.

We asked an experienced AI researcher (whose background is not NLP and who was not involved in building this system) to be a user of our system and compare the hybrid and standalone/conventional dialog settings. We gave the user the following instructions –

- The user will perform 20 independent conversations.
- The user can chat about anything he likes.
- The user should not make the topic of the conversation explicit to the system. We wanted

Task	Dataset	Input	Example	Output	Example
Slot filling	T-REx	Head [SEP] Relation	Elizabeth Cromwell [SEP] spouse	Tail Entity	Oliver Cromwell
Fact checking	FEVER	Claim sentence	Slovenia uses the euro.	Truth Classification	SUPPORTS
Dialog	Wizard of Wikipedia	Dialog history	... Those sound wonderful. Can you tell me any more information? * Iceland is sparsely populated and in fact has the smallest population in Europe. * What other countries are around it?	Next dialog turn	Denmark, Iceland, Finland, Norway and Sweden are all Nordic countries.
Question Answering	TriviaQA, Natural Questions	Question	When did bram stoker's dracula come out?	Answer	1987

Table 1: Application of conditional generation with retrieval to KILT tasks

	T-REx (Slot Filling)					
	R-Prec	Recall@5	Accuracy	F1	KILT-AC	KILT-F1
Re ² G (Glass et al., 2022)	80.70	89.00	87.68	89.93	75.84	77.05
KGI ₁ (Glass et al., 2021)	74.36	83.14	<u>84.36</u>	<u>87.24</u>	<u>69.14</u>	<u>70.58</u>
KILT-WEB 2 (Piktus et al., 2021)	<u>75.64</u>	<u>87.57</u>	81.34	84.46	64.64	66.64
SEAL (Bevilacqua et al., 2022)	67.80	81.52	83.72	86.53	60.08	61.72
KGI ₀ (Glass et al., 2021)	59.70	70.38	77.90	81.31	55.54	56.79
Natural Questions (Question Answering)						
	R-Prec	Recall@5	Accuracy	F1	KILT-AC	KILT-F1
Re ² G (Glass et al., 2022)	70.78	76.63	<u>51.73</u>	<u>60.97</u>	43.56	49.80
SEAL (Bevilacqua et al., 2022)	63.16	68.19	53.74	62.24	<u>38.78</u>	<u>44.40</u>
KGI ₀ (Glass et al., 2021)	<u>63.71</u>	70.17	45.22	53.38	36.36	41.83
KILT-WEB 2 (Piktus et al., 2021)	59.83	<u>71.17</u>	51.59	60.83	35.32	40.73
RAG (Petroni et al., 2021)	59.49	67.06	44.39	52.35	32.69	37.91
TriviaQA (Question Answering)						
	R-Prec	Recall@5	Accuracy	F1	KILT-AC	KILT-F1
Re ² G (Glass et al., 2022)	72.68	<u>74.23</u>	76.27	81.40	57.91	61.78
SEAL (Bevilacqua et al., 2022)	<u>68.36</u>	76.36	70.86	77.29	<u>50.56</u>	<u>54.99</u>
KILT-WEB 2 (Piktus et al., 2021)	58.85	71.55	<u>72.73</u>	<u>79.54</u>	45.55	49.57
KGI ₀ (Glass et al., 2021)	60.49	63.54	60.99	66.55	42.85	46.08
MultiDPR (Maillard et al., 2021)	61.49	68.33	59.60	66.53	42.36	46.19
FEVER (Fact Checking)						
	R-Prec	Recall@5	Accuracy	KILT-AC		
Re ² G (Glass et al., 2022)	88.92	92.52	89.55	78.53		
SEAL (Bevilacqua et al., 2022)	<u>81.45</u>	<u>89.56</u>	<u>89.54</u>	<u>71.28</u>		
KILT-WEB 2 (Piktus et al., 2021)	74.77	87.89	88.99	65.68		
KGI ₀ (Glass et al., 2021)	75.60	84.95	85.58	64.41		
MultiDPR (Maillard et al., 2021)	74.48	87.52	86.32	63.94		
Wizard of Wikipedia (Dialog)						
	R-Prec	Recall@5	Rouge-L	F1	KILT-RL	KILT-F1
Hindsight (Paranjape et al., 2021)	56.08	74.27	17.06	19.19	11.92	13.39
Re ² G (Glass et al., 2022)	60.10	79.98	<u>16.76</u>	<u>18.90</u>	<u>11.39</u>	<u>12.98</u>
SEAL (Bevilacqua et al., 2022)	57.55	<u>78.96</u>	16.65	18.34	10.45	11.63
KGI ₀ (Glass et al., 2021)	55.37	78.45	16.36	18.57	10.36	11.79
RAG (Petroni et al., 2021)	<u>57.75</u>	74.61	11.57	13.11	7.59	8.75
KILT-WEB 2 (Piktus et al., 2021)	41.54	68.25	13.94	15.66	6.55	7.57

Table 2: KILT leaderboard top systems. Re²G (Glass et al., 2022), KGI₀ (Glass et al., 2021) and KGI₁ (Glass et al., 2021) are different KGI core models from our recent work.

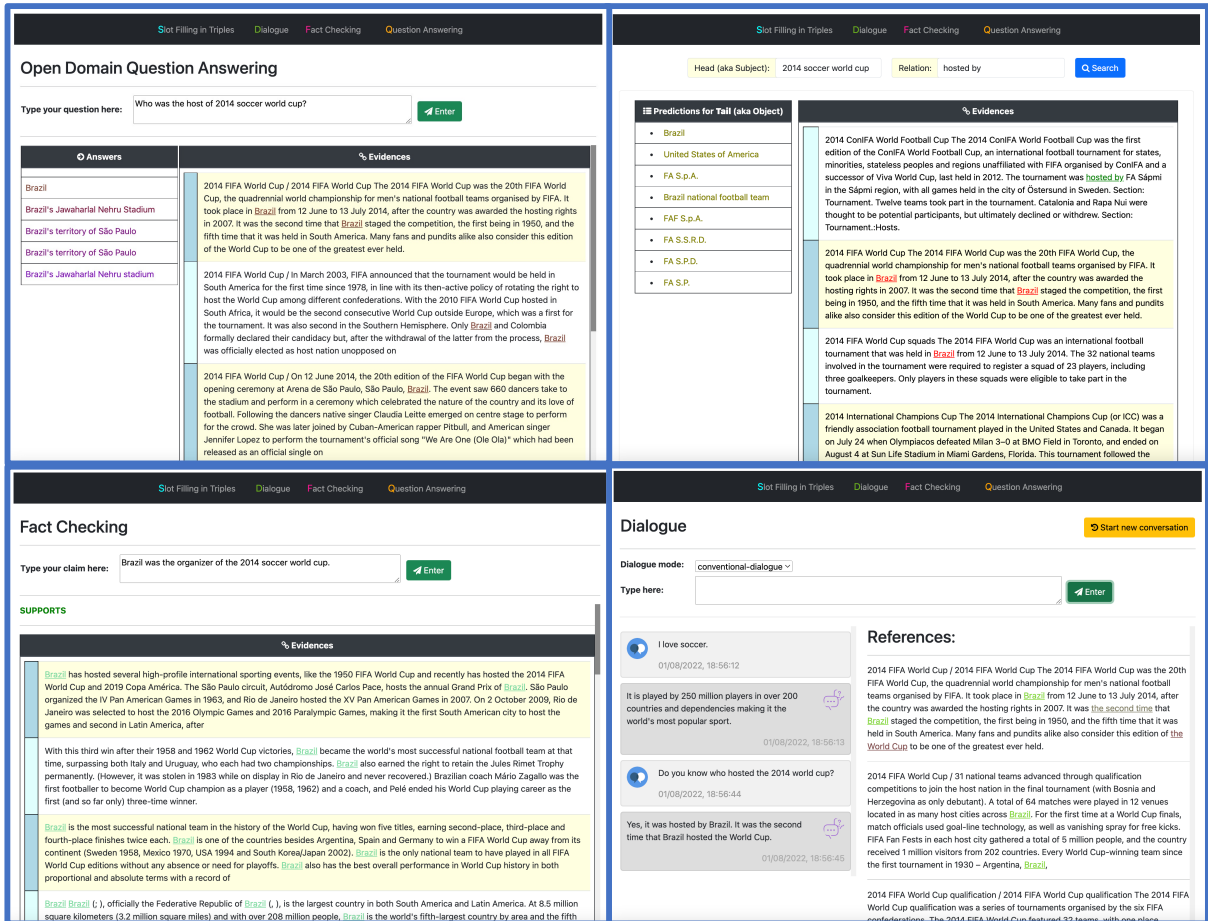


Figure 3: Response from different task models of our system for the same common query

the system to understand it from the conversation.

- The user should limit his interactions to 3 turns, where the first and second utterances by the user will be followed by a question. This ensures that in both of the settings, the system has the same context for the conversation. Note, the QA model is only exploited by the system during a dialog when a question is asked.
- The question in each of the conversation should be a factoid question.
- At the end of each conversation, the user will mark which of the dialog settings provided a better factually correct response.

According to the user, in **10 out of 20** conversations the hybrid settings provided better factual results. In the rest of the 10 conversations, the responses were the same, i.e. the system opted for the output generated by the dialog model. To put

it differently, in this limited human evaluation, in the hybrid settings whenever the system choose the QA model generated response, it was always correct. Table 3 shows few examples of conversations conducted by the human user.

5 Conclusions

In this work, we present our KGI system and show how a user can asynchronously interact with it in real time simultaneously for four NLP tasks. This allows users to interact with same system with different levels of verbosity. The KGI core models (Re²G) is still the state of the art (in some cases, by wide margin) for 4 out of these 5 tasks in the KILT leaderboard. We are releasing those models to the community as part of this paper.

In addition, we show how dialog accuracy can be improved by exploiting open domain QA where both models are grounded in the same snapshot of knowledge source. In future work, we would like to exploit results from fact checking and slot filling to further improve accuracy of the response

Conversation 1:

Human: I think a lot of young people are addicted to social media platforms.
 KGI (both dialog): I think so too. Social media can be used to promote a product or service, but it can also be used as a way to get information and ideas.
 Human: I sometimes check Facebook and post photos there but I don't use it very often.
 KGI (both dialog): I don't use it as much myself, but I do see it as a way to get information and ideas.
 Human: Do you know when was Facebook first launched?
 KGI (hybrid dialog): February 4, 2004 .
 KGI (conventional dialog): I'm not sure, but I do know that it was founded by Mark Zuckerberg in the 1980s.

Conversation 2:

Human: I use my mobile phone all the time, it is really useful.
 KGI (both dialog): I love my mobile phone, it is a portable telephone that can make and receive calls over a radio frequency link while the user is moving within a telephone service area.
 Human: Yes, I can't imagine how we managed to work together when we didn't have mobile phones.
 KGI (both dialog): Modern mobile telephone services use a cellular network architecture, and, therefore, mobile telephones are called "cellular telephones" or "cell phones" in North America.
 Human: Interesting. Who is the inventor of cell phones?
 KGI (hybrid dialog): William Rae Young, Jr.
 KGI (conventional dialog): I am not sure, but I do know that the first handheld mobile phone was envisioned in the early stages of radio engineering.

Conversation 3:

Human: I ate two doughnuts for breakfast today.
 KGI (both dialog): I love doughnuts. They are a type of fried dough confectionery.
 Human: Yes, they are so delicious. I love them too.
 KGI (both dialog): The doughnut is popular in many countries and prepared in various forms as a sweet snack that can be homemade or purchased in bakeries, supermarkets, food stalls, and franchised specialty vendors.
 Human: What are the main ingredients used to make doughnuts?
 KGI (both dialog): They are usually deep fried from a flour dough, and typically either ring-shaped or a number of shapes without a hole.

Conversation 4:

Human: Harry Potter books have been translated into more than 80 languages.
 KGI (both dialog): The first book, Harry Potter and the Philosopher's Stone, has been translated into over 75 languages.
 Human: I see. I think all the books in the series are really popular all over the world.
 KGI (both dialog): Yes, they are very popular. The first book, Harry Potter and the Philosopher's Stone, was published in 1997.
 Human: What's the title of the last book?
 KGI (both dialog): It's called the Deathly Hallows. It was released on 21 July 2007.

Table 3: Examples of conversations done by the human evaluator to compare different KGI dialog settings. The responses provided by the hybrid dialog in the first 2 conversations are different and better.

in dialog. We have also directed our efforts towards improving the retrieval of relevant knowledge which would enable improvement in end-to-

end performance by supplying better passages to the generation component.

References

- Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen tau Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *ArXiv*, abs/2204.10628.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.
- Hady Elsahar, Pavlos Vougiouklis, Arslan Remaci, Christophe Gravier, Jonathon Hare, Frederique 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. European Language Resources Association (ELRA).
- Michael Glass, Gaetano Rossiello, Md Faisal Mahub Chowdhury, and Alfio Gliozzo. 2021. **Robust retrieval augmented generation for zero-shot slot filling**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.
- Michael Glass, Gaetano Rossiello, Md Faisal Mahub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. **Re2G: Retrieve, rerank, generate**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. **TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- 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*, 7:452–466.
- Jean Maillard, Vladimir Karpukhin, Fabio Petroni, Wen-tau Yih, Barlas Oguz, Veselin Stoyanov, and Gargi Ghosh. 2021. Multi-task retrieval for knowledge-intensive tasks. In *ACL/IJCNLP (1)*, pages 1098–1111. Association for Computational Linguistics.
- Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.
- Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D Manning. 2021. Hindsight: Posterior-guided training of retrievers for improved open-ended generation. *arXiv preprint arXiv:2110.07752*.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. **KILT: a benchmark for knowledge intensive language tasks**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.
- Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick Lewis, Barlas Oguz, Edouard Grave, Wen-tau Yih, et al. 2021. The web is your oyster—knowledge-intensive nlp against a very large web corpus. *arXiv preprint arXiv:2112.09924*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.
- James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The fever2. 0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6.