Exploring Argument Retrieval with Transformers

Notebook for the Touché Lab on Argument Retrieval at CLEF 2020

Christopher Akiki and Martin Potthast

Leipzig University

Abstract We report on our recent efforts to employ transformer-based models as part of an information retrieval pipeline, using argument retrieval as a benchmark. Transformer models, both causal and bidirectional, are independently used to expand queries using generative approaches as well as to densely embed and retrieve arguments. In particular, we investigate three approaches: (1) query expansion using GPT-2, (2) query expansion using BERT, and orthogonal to these approaches, (3) embedding of documents using Google's BERT-like universal sentence encoder (USE) combined with a subsequent retrieval step based on a nearest-neighbor search in the embedding space. A comparative evaluation of our approaches at the Touché lab on argument retrieval places our query expansion based on GPT-2 first on the leaderboard with a retrieval performance of 0.808 nDCG@5, improving over the task baseline by 6.878%.

1 Introduction

Search has become the sine qua non tool of information access, and the gateway to the World Wide Web. The users of web search engines meanwhile expect a high quality of the search results in terms of their relevance to the queries submitted: If relevant documents exist for a given query, they are usually found, and the most relevant ones can be expected to be ranked highest. However, search engines are optimized for ad hoc retrieval tasks, and a key assumption is that a single document suffices to satisfy the information need underlying a query. That assumption falls apart when the topic of interest is inherently subjective and nuanced, such as is the case for contentious issues. Any one document will most likely argue one way or another, so that a user may need to peruse many documents at a time to satisfy a deliberative information need. The current search landscape is, however, not especially attuned to such a nuance, usually preferring to let "the stakeholders compete for what opinion ranks higher" [23]. The ability to specifically handle arguments rather than the documents that might contain them is an attempt to address that problem using computational argumentation analysis [23, 31]. Such an approach forms the basis of the args.me search engine which relies on a corpus of more than 300,000 arguments mined from online debate portals [1]. This corpus formed the basis of the Touché shared task on argument retrieval [3].

We leverage transformer models to see how they might enrich argument retrieval at various stages of the information retrieval (IR) pipeline. Attention-based transformer models [30] have seen a recent surge in popularity as their readiness for massive parallelization and the increasing availability of computational power on GPUs led to such models claiming the state-of-the-art crown on many natural language processing and understanding tasks [7, 25]. We show that competitive performance can be achieved on this task with no preprocessing of the documents or fine-tuning of the models on the task. In what follows, Section 2 presents related work, Section 3 exposes our approach, and Section 4 presents the results of our participation in the shared task.

2 Related Work

Although the subject of arguments in a retrieval context is by no means a new development (see, e.g., [26] and [28]), the field itself is still nascent. Lawrence and Reed [18], Potthast et al. [23], and Wachsmuth et al. [31] provide a comprehensive overview of the field, with the last two further putting forth the theoretical considerations that underlie the development of a tool where arguments form the unit of retrieval, using the args.me dataset [1] to ground the theory in an applied retrieval study. Another salient research direction is spearheaded by IBM's Project Debater and corresponding studies that leverage its numerous datasets (see [29] for an overview of the project).

The growing interest of natural language processing (NLP) researchers in information retrieval and argument mining [27], as well as the impressive performance and ease of use of transformer-based models [32] and their propensity to semantic nuance makes a convergence of the two unavoidable. Indeed, researchers in IBM's project debater have recently published two corresponding papers [9, 14] that both use BERT [7] for argument retrieval and argument quality assessment. Like other efforts in the argument mining direction [6, 11], these methods, though impressive in their results, do not readily transfer to a domain other than that of the retrieval of arguments because they are usually fine-tuned on a specific document corpus.

The use of transformers in general, specifically BERT, to the field of document retrieval was until very recently limited to frameworks where initial retrieval is delegated to the Lucene-based Anserini retrieval toolkit [34, 33], which, while proving a promising approach, did not attempt to instrumentalize transformers at different parts of the IR pipeline. Another approach similarly leverages BERT and Anserini for ad hoc document retrieval, while also coupling the approach with the ability to interface with Python to simplify neural network based research [35].

The semantic prowess of transformers makes them prime candidates for enriching IR pipelines that rely on query expansion [2]. For a deeper coverage of query expansion we refer the reader to Azad and Deepak's [2] thorough survey on the topic. In brief, query expansion consists in augmenting a user query to increase its effectiveness and reduce ambiguity. That is achieved through reformulating the query using additional data, and the source of that data coincides with the different sub-approaches of this sub-field of IR. Azad and Deepak [2] differentiate between two main directions: global analysis and local analysis. The latter relies on user feedback, both direct and implicit, whereas the former consists of approaches that rely on knowledge that can be gleaned

and derived from the query itself or its context. These include linguistic, corpus-based, log-based and web-based approaches. Global analysis has proven to be of particular interest for transformer-focused research [8, 22, 19, 21]. The work of Dibia [8] in particular aligns nicely with our approach in Section 3.2. Our approaches differ however in the strategies used to determine where and how to inject context. To our knowledge, the query expansion approach we develop in Section 3.1 is the first use of a transformer decoder (GPT-2 in this instance) to generate documents that read as though they might have originated from a corpus of interest, at least plausibly enough for a retrieval system, and thus narrow down the scope of search.

3 Using Transformers for Document Retrieval

We set out to instrumentalize for the retrieval of documents the ability of different transformers to encode knowledge. The original transformer introduced by Vaswani et al. [30] uses a standard seq2seq encoder-decoder architecture, whereby the encoder learns a task-specific embedding, and the decoder learns a language model. Subsequent transformer-based models do not necessarily follow that convention. BERT [7] only uses an encoder, and GPT-2 [25] only a decoder. It is therefore useful to qualify the rather general "transformer" nomenclature with either "encoder" or "decoder".

We use transformer decoders (i.e., GPT-2-like models) for query expansion via text hallucination (Section 3.1), i.e., the generation of a text that reads as if it might have come from the corpus, and transformer encoders (i.e., BERT-liked models) for keyword-based query expansion (Section 3.2). Moreover, we consider transformer encoders for document embedding (Section 3.3). Both query expansion approaches make use of an Elasticsearch index where the documents of the args.me corpus were indexed using a language model with Dirichlet priors (DirichletLM) [36], which has been shown to be superior to other retrieval models for retrieving arguments from the args.me corpus [23]. The embedding approach, on the other hand, uses a vector-based similarity index for retrieval.

Self-supervised pre-training on massive amounts of qualitatively diverse data is what enables transformer models to encode the knowledge that allows them to perform as well as they do on natural language processing (NLP) and natural language understanding (NLU) tasks. Therein also lies their promise for retrieval tasks. We make the conscious decision to only rely on the knowledge encoded into the models by these tasks; that is, we do not rely on any argumentation-specific fine-tuning. That allows us to gauge the performance of transformers for retrieval tasks in general. In particular, our investigation aims at a modular proof-of-concept approach, not to show the superiority of a certain method over others, nor was it our goal to optimize the ensuing pipeline for accuracy on the relevance judgments.

3.1 Query Expansion with a Transformer Decoder

A lot of the recent media hype around transformer networks is centered around their ability to generate coherent text. Transformer decoders can be trained as causal language models (CLM), seeing as the representation of a given token can only depend

Table 1. Language model prompts for argumentative query expansion. The dots (. . .) indicate where the language model takes over and generates text that is used as expanded query.

Positive Prompt	Negative Prompt	Neutral Prompt
- What do you think? **<query>** - Yes because . . .	- What do you think? **<query>** - No because . . .	- What do you think? **<query>** - I don't know . . .
- What do you think? **<query>** - The answer is yes . . .	- What do you think? **<query>** - The answer is no . . .	- What do you think? **<query>** - Not sure . . .

on the tokens preceding it. Indeed, since transformer decoders like GPT-2 simulate language models, it is possible to iteratively generate sentences by sampling from the output distribution at a given time step and feeding that output back into the network at the following time step. By choosing an opening for a text (called prompt), it becomes thus possible to steer GPT-2 and exert some influence over the text it generates. Our approach for query expansion makes use of this capability to generate argumentative expansions of a given query with a positive, negative, and a neutral stance by prompting GPT-2 in turn with the six prompts shown in Table 1. The prompts are purposefully constructed so as to simulate an argumentative dialog that is to be completed by GPT-2.

The quality of generated sequences, and thus also of our retrieval results, is highly contingent upon the way in which tokens are sampled from the network. It is easy for neural language models to degenerate into incoherent, non-human sounding text when using naive likelihood maximization [16]. Many sampling methods exist that try to make sure the generated text is less likely to be incoherent, repetitive, or overly generic [12, 13, 16]; the following sampling approaches are the most salient ones, ordered from basic to advanced:

- *Pure sampling* is the most naive as well as often the worst-performing sampling method. It consists of greedily choosing the most likely token at every time step. This sampling method usually leads to low quality text.
- *Beam Search* is a heuristic that keeps a set number of candidate sequences until they all reach the desired length and keeps the most likely candidate. Search for the output sequence that actually maximizes likelihood is intractable, so that beam search provides a reasonable alternative in practice.
- *Temperature scaling* reshapes the output distribution by skewing it either toward high probability tokens or low probability tokens: the former improves generation quality but hurts diversity [16], while the latter regularizes generation by making the model less certain of its top choices.
- *Top-k sampling* is the sampling method that was used for GPT-2 [25]. It truncates the language model's output distribution to a set of size k of most likely tokens. The sampling then uses relative probability within this truncated set. All other tokens are not considered for generation.
- *Nucleus sampling* is a stochastic decoding method proposed by Holtzman et al. [16]. Similar to top-k sampling, it also truncates the output distribution, albeit by setting a threshold the cumulative distribution of top tokens must reach before discarding the rest of the tokens.

For each one of the six prompts in Table 1, we use GPT-2 to generate four possible continuations up to a maximum of 100 tokens. To ensure the four continuations are different from each other, we chose different combinations of the aforementioned sampling strategies.[1] Having framed the prompts as conversational (by virtue of the dashes and formatting), the ensuing generated text often tends to read like an argument. Having generated 24 such texts,[2] we discard the original query and use these hallucinations as 24 queries against the DirichletLM index, which we combine additively to generate the final rankings. We only consider those documents which were returned by at least twelve of the queries simultaneously.[3]

For illustration purposes, consider the query *"Can alternative energy effectively replace fossil fuels?"*. Three of the 24 texts that our approach generates are:

- Yes, because *it has proven to be a significant and lasting improvement in fuel efficiency, carbon neutrality. The only other thing that could possibly help this energy is the need for nuclear reactors at low cost which would require more than 20 percent of current generation electricity by 2030 (currently under construction), plus less renewable resources like wind or solar power as well but with sufficient amounts of coal/solar panels if there are enough* [sic]
- No, because *there is no evidence for that. There are a few possible alternative energy options available to people who would like to cut down fossil fuels, and I believe those include wind or solar power."The main thing we want in our future climate policies has to be better use of resources instead on these things than if they weren't used at the moment," said Kieferlein. "We need clean air* [sic]
- Not sure. *However, many scientists have made the point that alternative sources of power are already producing more carbon emissions than they would otherwise (and it seems like such a small number in our country). There has been some debate about whether this was actually true or if there simply wasn't much coal available at all to replace fossil fuels and other forms thereof as an environmentally sustainable form... In fact, recent studies suggest we* [sic]

3.2 Query Expansion with a Transformer Encoder

Unlike transformer decoders, transformer encoders like BERT [7] cannot be used to auto-regressively generate sentences. Their attention-based architecture is such that any token can see ("attend to") any other token of the sentence. Being able to look ahead into the future of a sequence breaks the causality required by a CLM. As such, these models have to rely on other pre-training tasks to gain linguistic coherence. Global linguistic coherence is achieved through next sentence prediction, where BERT has to predict whether two sentences follow each other in a corpus. Local coherence is achieved through another pre-training task, namely masked language modeling (MLM).

[1] One with greedy sampling using 10 beams, and three using a temperature of 1.6, a top-k threshhold of 100 tokens, and a nucleus sampling probability threshhold of 0.4.

[2] Six prompts with four continuations per prompt.

[3] This corresponds to an Elasticsearch boolean query of type "should" with the min_should_match parameter set to 12.

Table 2. Masked language model prompts for argumentative query expansion. The masks (`[MASK]`) indicate where the language model takes over and generates alternative words that are used as expanded query.

Stance	Prompt
Positive	- What do you think? **<query>** - Yes, because of **[MASK]** and the benefits of **[MASK]**.
	- What do you think? **<query>** - Absolutely, I think **[MASK]** is good!
	- What do you think? **<query>** - Yes, **[MASK]** is associated with **[MASK]** during **[MASK]**.
Negative	- What do you think? **<query>** - No, because of **[MASK]** and the risk of **[MASK] [MASK]**.
	- What do you think? **<query>** - Absolutely not, I think **[MASK]** is bad!
	- What do you think? **<query>** - No, **[MASK]** is associated with **[MASK]** during **[MASK]**.
Neutral	- What do you think? **<query>** - What about **[MASK]** or **[MASK]**?
	- What do you think? **<query>** - Don't forget about **[MASK]**!

During training, tokens in the input are masked at random and the network is tasked with guessing what that word was. Learning to fill in the blank has been shown to imporve the quality of text generation [10].

We leverage this ability of the model to "fill in the blank" to enrich the original query with a set of words that are contextually relevant to the topic at hand. To achieve that, we again augment the original query using the same strategy as outlined in the previous section, this time, however, we leave blanks for BERT to fill out in the form of the `[MASK]` token (see Table 2). For every `[MASK]` in every augmented seed text, we ask BERT to return the five most likely words, filtering out stop words, punctuation, and sub-words. This amounts to an average (min=206, max=473) of 340 thematic keywords per query. All keywords are then joined together into a space-separated list of keywords which is what we use to query the DirichletLM index, discarding the original query. For illustration, consider the query *"Can Alternative Energy Effectively Replace Fossil Fuels?"* and provide the resulting keywords when expanding the query with BERT:

> *diesel, cost, nuclear, consumption, hydrogen, technologies, energy, future, electricity, pregnancy, coal, alternative, migration, emissions, efficiency, economics, technology, growth, wartime, earthquakes, green, environmental, accidents, costs, renewable, winter, development, pollution, new, stress, water, oil, accident, death, health, warming, sustainability, accidental, fires, competition*

3.3 Document Embedding with a Transformer Encoder

Our final approach also employs a transformer encoder in the form of the large variant of Google's universal sentence encoder (USE) [5]. Though architecturally similar to

BERT, the USE model is trained with very different pre-training tasks, which were specifically picked to perform well on downstream NLU tasks with the goal of creating a sentence embedder. A further distinction to BERT is USE's unbounded input length, which lends itself well to the args.me corpus. The following pre-training tasks were considered:

- *Skip-thought* is a self-supervised pretraining task, originally devised to use LSTMs to provide high-quality sentence vectors by training on a large amount of contiguous text [17].
- *Natural language response suggestion* imparts conversational awareness to the sentence encoder, which fits quite well to the task at hand. The goal of this supervised task is to predict the best short reply among millions of options in response to an email [15].
- *Stanford natural language inference* is a labeled dataset [4] of 570,000 sentence pairs. This can be seen as a supervised variant of BERT's next sentence prediction pre-training task. In this instance however, entailment, contradiction, and irrelevance are explicitly labeled in the data itself, rather than implied by the relative position of two sentences in an unlabeled corpus of contiguous text.

These pre-training tasks make USE a great candidate for argument retrieval. Using USE, we embed each document in the args.me corpus into a 512-dimensional space. To retrieve arguments given a query, we embed the query using the same model into the same space, and perform exhaustive nearest-neighbor search,[4] considering both L2 distance and inner-product (IP) distance for retrieval.

We carried out a small pilot experiment to make sure USE projects the args.me corpus in a semantically meaningful way by running k-means on the embedded corpus, choosing a cluster size of 100. The clusters obtained are both syntactically and semantically coherent in a way that is surprisingly meaningful. Some clusters are thematically coherent, encompassing topics, such as religion, politics, and economics. Others are both syntactically and semantically coherent, where all premises are of the form "X is better than Y" and covering themes such as video game consoles, superheroes, and consumer electronics. Further clusters are only syntactically coherent, where, for instance, all arguments consist of YouTube links or of repeated short idiosyncratic phrases one tends to find on online debate websites (e.g., "I agree.").

4 Evaluation

The evaluation of our approaches to argument retrieval was carried out as part of the Touché shared task. In what follows, we briefly recap the experimental setup and overview the performance achieved.

4.1 Experimental Setup

The Touché shared task on argument retrieval [3] uses the TIRA evaluation platform [24] to judge entries to the competition. On TIRA every task participant is assigned their own virtual machine, and submitting a retrieval model to the shared task

[4] https://github.com/facebookresearch/faiss

Table 3. Evaluation results of our approaches compared to the two runner-ups of the shared task.

Model	nDCG@5	nDCG@10	nDCG	QrelCoverage@10
GPT-2	**0.808**	0.586	0.378	5.70
Baseline (DirichletLM)	0.756	–	–	–
BERT	0.755	0.538	0.337	5.36
Team Aragorn	0.684	–	–	–
USE (L2)	0.598	0.397	0.285	4.16
Team Zorro	0.573	–	–	–
USE (IP)	0.527	0.36	0.275	3.82

corresponds to submitting software to be run on that virtual machine with the relevant inputs provided by TIRA at run time. Those inputs include the args.me corpus and a list of 50 topics (queries) on which the retrieval model is to be judged using crowdsourced relevance judgments. Though participant entries are ranked by nDCG@5 scores on the leaderboard,[5] TIRA also returns nDCG@10, nDCG, and QrelCoverage@10, which we include in Table 3.

4.2 Results

The results of all our runs are included in Table 3, where one can clearly see the considerable improvement over the official baseline by our GPT-2 query expansion approach, which comes out on top of the leaderboard. Our BERT query expansion model basically ties with the baseline of 0.756 as it manages to score an nDCG@5 score of 0.755. We speculate that this performance might be partly due to the fact that both the args.me corpus and the datasets on which BERT and GPT-2 are trained consist of user-generated internet data. Both embedding-based runs perform worse than the query expansion approaches and that is to be expected, as the only information signal afforded to those runs originates in the query itself, whereas the other approaches had the benefit of considerable added context through query expansion. Still, judging the embeddings on their own constitutes a useful baseline. A promising approach would combine these two orthogonal approaches.

5 Conclusion

This work showcases three possible uses of transformer models for the retrieval of relevant arguments from the args.me corpus in particular, and document retrieval from a corpus in general. Impressive results were achieved without hyperparameter tuning or optimization of any sort. A promising future continuation of this work would be an ablation study that judges the effect that the hyperparameters have on retrieval-optimized text generation. Such a continuation would be necessary to judge the quantitative merits of each approach.

[5] https://events.webis.de/touche-20/shared-task-1.html

It is also important to note that BERT and GPT-2 are merely the representatives of the transformer family of models. There exists a myriad of models[6] that build on the foundations laid down by the works that introduced them, iterating, improving, and filling gaps those original models did not take into account. It would therefore be important to experiment with other models, perhaps also come up with new pre-training tasks that would make query expansion even more performant.

Furthermore, our approach leaves out any sort of natural language preprocessing of the corpus or fine-tuning of any of the used models. That some approaches perform as well as they do is a testament to the amount of linguistic and world knowledge encoded in the weights of pre-trained transformers. A future research direction might leverage the args.me and Project Debater corpora to add more argumentative awareness to the transformers and indubitably improve retrieval results.

Finally, we believe it crucial to sensibly modulate any research direction with the due ethical considerations of such projects. It is unclear whether to include user feedback, as including such signals would incur the risk of calcifying existing biases. While it might be useful to think of a search engine as an educational tool, it might prove dangerous to assume it is the prerogative of an information retrieval technology to monopolize the task of teaching its users how to think by conditioning them to blindly rely on it to populate their existing biases.

Bibliography

[1] Ajjour, Y., Wachsmuth, H., Kiesel, J., Potthast, M., Hagen, M., Stein, B.: Data Acquisition for Argument Search: The args.me corpus. In: Benzmüller, C., Stuckenschmidt, H. (eds.) 42nd German Conference on Artificial Intelligence (KI 2019), pp. 48–59, Springer, Berlin Heidelberg New York (Sep 2019), https://doi.org/10.1007/978-3-030-30179-8_4

[2] Azad, H.K., Deepak, A.: Query expansion techniques for information retrieval: A survey. Information Processing & Management **56**(5), 1698?1735 (Sep 2019), ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2019.05.009, URL http://dx.doi.org/10.1016/j.ipm.2019.05.009

[3] Bondarenko, A., Fröbe, M., Beloucif, M., Gienapp, L., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2020: Argument Retrieval. In: Working Notes Papers of the CLEF 2020 Evaluation Labs (Sep 2020), ISSN 1613-0073

[4] Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y. (eds.) Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, pp. 632–642, The Association for Computational Linguistics (2015), https://doi.org/10.18653/v1/d15-1075, URL https://doi.org/10.18653/v1/d15-1075

[5] Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., Kurzweil, R.: Universal sentence encoder. CoRR **abs/1803.11175** (2018), URL http://arxiv.org/abs/1803.11175

[6] Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., Hwang, A.: AMPERSAND: argument mining for persuasive online discussions. In: Inui, K., Jiang, J., Ng, V., Wan, X.

[6] https://github.com/thunlp/PLMpapers

(eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pp. 2933–2943, Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/D19-1291, URL https://doi.org/10.18653/v1/D19-1291

[7] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), URL http://arxiv.org/abs/1810.04805

[8] Dibia, V.: Neuralqa: A usable library for question answering (contextual query expansion + bert) on large datasets (2020)

[9] Ein-Dor, L., Shnarch, E., Dankin, L., Halfon, A., Sznajder, B., Gera, A., Alzate, C., Gleize, M., Choshen, L., Hou, Y., Bilu, Y., Aharonov, R., Slonim, N.: Corpus wide argument mining - A working solution. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pp. 7683–7691, AAAI Press (2020), URL https://aaai.org/ojs/index.php/AAAI/article/view/6270

[10] Fedus, W., Goodfellow, I.J., Dai, A.M.: Maskgan: Better text generation via filling in the _____. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net (2018), URL https://openreview.net/forum?id=ByOExmWAb

[11] Fromm, M., Faerman, E., Seidl, T.: TACAM: topic and context aware argument mining. In: Barnaghi, P.M., Gottlob, G., Manolopoulos, Y., Tzouramanis, T., Vakali, A. (eds.) 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019, Thessaloniki, Greece, October 14-17, 2019, pp. 99–106, ACM (2019), https://doi.org/10.1145/3350546.3352506, URL https://doi.org/10.1145/3350546.3352506

[12] Géron, A.: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media (2019)

[13] Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT press Cambridge (2016), ISBN 978-0-262-03561-3

[14] Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., Slonim, N.: A large-scale dataset for argument quality ranking: Construction and analysis. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pp. 7805–7813, AAAI Press (2020), URL https://aaai.org/ojs/index.php/AAAI/article/view/6285

[15] Henderson, M.L., Al-Rfou, R., Strope, B., Sung, Y., Lukács, L., Guo, R., Kumar, S., Miklos, B., Kurzweil, R.: Efficient natural language response suggestion for smart reply. CoRR **abs/1705.00652** (2017), URL http://arxiv.org/abs/1705.00652

[16] Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net (2020), URL https://openreview.net/forum?id=rygGQyrFvH

[17] Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pp. 3294–3302 (2015), URL http://papers.nips.cc/paper/5950-skip-thought-vectors

[18] Lawrence, J., Reed, C.: Argument mining: A survey. Computational Linguistics **45**(4), 765–818 (2020), https://doi.org/10.1162/coli_a_00364, URL https://doi.org/10.1162/coli_a_00364
[19] Lin, S.C., Yang, J.H., Nogueira, R., Tsai, M.F., Wang, C.J., Lin, J.: Query reformulation using query history for passage retrieval in conversational search (2020)
[20] Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
[21] Naseri, S., Dalton, J.S., Allan, J., Yates, A.: Ceqe: Contextualized embeddings forquery expansion (2020)
[22] Padaki, R., Dai, Z., Callan, J.: Rethinking query expansion for bert reranking. In: European Conference on Information Retrieval, pp. 297–304, Springer (2020)
[23] Potthast, M., Gienapp, L., Euchner, F., Heilenkötter, N., Weidmann, N., Wachsmuth, H., Stein, B., Hagen, M.: Argument search: Assessing argument relevance. In: Piwowarski, B., Chevalier, M., Gaussier, É., Maarek, Y., Nie, J., Scholer, F. (eds.) Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, pp. 1117–1120, ACM (2019), https://doi.org/10.1145/3331184.3331327, URL https://doi.org/10.1145/3331184.3331327
[24] Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA integrated research architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, The Information Retrieval Series, vol. 41, pp. 123–160, Springer (2019), https://doi.org/10.1007/978-3-030-22948-1_5, URL https://doi.org/10.1007/978-3-030-22948-1_5
[25] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
[26] Rahwan, I., Zablith, F., Reed, C.: Laying the foundations for a world wide argument web. Artificial Intelligence **171**(10), 897–921 (2007), ISSN 0004-3702, https://doi.org/https://doi.org/10.1016/j.artint.2007.04.015
[27] Reed, C. (ed.): Proceedings of the Third Workshop on Argument Mining (ArgMining2016), Association for Computational Linguistics, Berlin, Germany (Aug 2016), https://doi.org/10.18653/v1/W16-28, URL https://www.aclweb.org/anthology/W16-2800
[28] Teufel, S., et al.: Argumentative zoning: Information extraction from scientific text. Ph.D. thesis, Citeseer (1999)
[29] Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., Slonim, N.: Automatic argument quality assessment - new datasets and methods. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pp. 5624–5634, Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/D19-1564, URL https://doi.org/10.18653/v1/D19-1564
[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 5998–6008 (2017), URL http://papers.nips.cc/paper/7181-attention-is-all-you-need
[31] Wachsmuth, H., Potthast, M., Khatib, K.A., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J., Stein, B.: Building an argument search engine for the web. In: Habernal, I., Gurevych, I., Ashley, K.D., Cardie, C., Green, N., Litman, D.J., Petasis, G., Reed, C., Slonim, N., Walker, V.R. (eds.) Proceedings of the 4th Workshop on Argument

Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017, pp. 49–59, Association for Computational Linguistics (2017), https://doi.org/10.18653/v1/w17-5106, URL https://doi.org/10.18653/v1/w17-5106

[32] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing. CoRR **abs/1910.03771** (2019), URL http://arxiv.org/abs/1910.03771

[33] Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., Li, M., Lin, J.: End-to-end open-domain question answering with bertserini. In: Ammar, W., Louis, A., Mostafazadeh, N. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations, pp. 72–77, Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/n19-4013, URL https://doi.org/10.18653/v1/n19-4013

[34] Yang, W., Zhang, H., Lin, J.: Simple applications of BERT for ad hoc document retrieval. CoRR **abs/1903.10972** (2019), URL http://arxiv.org/abs/1903.10972

[35] Yilmaz, Z.A., Wang, S., Yang, W., Zhang, H., Lin, J.: Applying BERT to document retrieval with birch. In: Padó, S., Huang, R. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations, pp. 19–24, Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/D19-3004, URL https://doi.org/10.18653/v1/D19-3004

[36] Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 334–342, SIGIR '01, Association for Computing Machinery, New York, NY, USA (2001), ISBN 1581133316, https://doi.org/10.1145/383952.384019, URL https://doi.org/10.1145/383952.384019