Question Answering for Comparative Questions with GPT-2

Notebook for Touché at CLEF 2020

Bjarne Sievers[1]

University of Leipzig, Germany
b.sievers@studserv.uni-leipzig.de

Abstract Finding the best answer for comparative questions is a difficult problem in the field of Information Retrieval. Since the best answer ideally covers not only both subjects of the query, but puts them into relation with each other, keyword based search alone has difficulties understanding the users intention correctly. Language models like GPT-2 on the other hand can distinguish fine nuances of intention in a query or sentence and generating a text conditioned on that phrase. We try to leverage this by generating substitute answer articles conditioned on the given query. Search results are then re-ranked by their textual similarity to the generated substitute answer articles.

1 Introduction

Nowadays, search engines are not only used for looking up pure facts, but also for opinions, discussions and arguments. Users now turn to search engines with questions like ”Which is healthiest: coffee, green tea or black tea and why?” or ”Do you prefer tampons or pads?”. In these cases, the user is likely more interested in the arguments or personal testimonies than in the final answer. Platforms like Quora often list the exact question and several answers to it, but depending on the answer, little context might be given. Discussion forums or comments on private blogs might contain more interesting debates and replies.

However, search engines still struggle with this type of questions. They usually either return websites that list facts for parts of the query or websites where the exact same question was posted. If a question was phrased in a different way, most search engines have trouble returning good results.

Language models like GPT-2 on the other hand have proven to precisely capture the semantics of a given prompt. Conditioned on such a prompt, they can generate almost arbitrary lengths of coherent text while understanding what the prompt is about. This has been demonstrated to be useful for summarization, translation and question answering[13].

The second task of the ”1st Shared Task on Argument Retrieval” in Touché@CLEF 2020 [4] is about returning the best results for comparative questions from the ClueWeb12

Corpus [1]. For this task, we propose using a language model like GPT-2 to generate the hypothetically perfect answer to a question. We can then compare this answer with actual search results from a classical search engine. Re-ranking them by their similarity with the generated answer, we produce search results that are more useful.

2 Related Work

2.1 Argument Retrieval

Argument Retrieval is a part of Information Retrieval that is focused on providing the best arguments, either for debates of for comparative questions. Most research to date is concerned with mining arguments from the web, identifying argument units [6] and making them searchable [18]. The Comparative Argumentative Machine [15] can search the web for comparisons between two entities and summarize the tendency of online publications towards either of them in percent.

2.2 Deep Learning

Since the success of Deep Learning with Convolutional Neural Networks [10] at the ImageNet competitions, Deep Learning has been used very successfully in many areas. In the field of speech processing, Recurrent Neural Networks (RNNS) are used, for example LSTMs [9]. Long dependencies that exist in natural language can be a challenge here: words can refer to past phrases or sentences that are even further back in the sequence. To counter this problem, approaches are often used that can dynamically decide on which previous tokens they focus their attention [17]. These transformers combine CNNs with a modification of attention, the so-called self-attention. Transformers thus form the basis of almost all current generative language models, for example GPT-2 [13] and T-NLG [2].

2.3 Text Similarity

There are different approaches to capture the similarity between texts. [8] divides the similarity metrics into three main categories: Similarity at the character or word level, similarity by comparing with large amounts of text, and semantic similarity calculated by semantic word graphs.[14] is a Neuronal Network pretrained on NSLI [5] and MultiSLI [19] on the basis of BERT [7]. It generates semantic vectors at sentence level that can be compared to each other, for example, via their cosine similarity.

3 Approach

The architecture used in this approach consists of three steps. First, multiple queries are generated from the given query by replacing words with synonyms, one by one. Second, results for all queries are queried with ChatNoir [3] and accumulated. Third, the retrieved results are re-ranked by their similarity with GPT-2 generated texts conditioned on the original query.

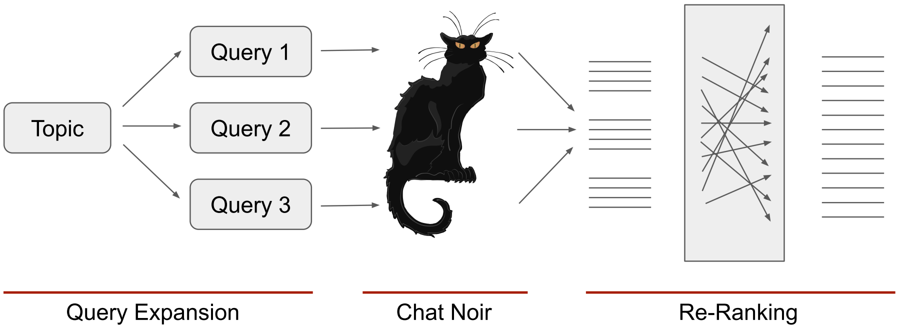


Figure 1: Architecture of the proposed approach.

3.1 Language Model

In order to best satisfy the information deficit of the users of our system, we have to provide an answer that is precisely tailored to the question. Lets suppose we understand exactly what the information deficit is and we have the information and knowledge to give the answer directly. And even if we do not have this knowledge, we would probably use very similar words and expressions in our answer, like a document representing the perfect answer. A system which formulates an answer for a given search query which at first glance does not differ significantly from a human answer, thus offers the possibility to reduce the problem finding the best answer to a classical information retrieval problem. The interpretation of which aspect of the question was meant by the user and what answer style they are expecting is already decided by this system. The generated mock document can now be compared with the corpus (here CW12[1]) by classical search engines and existing documents can be used as an answer.

Such systems have become conceivable due to the success of Deep Learning (DL) in Natural Language Processing (NLP). GPT-2 [13] from OpenAI or T-NLG from Microsoft prove that DL systems are already able to distinguish fine nuances of meaning even in short text snippets and to generate and summarize suitable texts or even answer questions. Our approach uses the GPT-2 model to generate conditioned query mock responses on the query. These generated texts are then used to find matching, similar documents in the corpus. We show that by using suitable suffixes on the question, certain answer formats can be favored and thus influence the results.

In order to find personal opinions, field reports, subjective impressions of the question, one can, for example, add generic sentences to the question that one would read in forums or question-answer platforms. If, on the other hand, you are more interested in a scientific approach to the text or in journalistic texts, you can use appropriate phrases. A corresponding overview of tested suffixes can be found in 1.

3.2 Implementation[1]

Before querying ChatNoir with the topic, each query is duplicated and adapted several times. This is done by removing all stopwords from the query and generating n queries, where n is the number of remaining words. In each query, one of the words is now substituted by its closest neighbor in the GloVe-graph[12] that has a Levenshtein distance greater than three. The Levenshtein distance is used in order to retrieve words that are neither singular or plural forms nor spelling mistakes of the original word. ChatNoir is now queried with each query and the results are downloaded and accumulated.

For each original query, the system uses the GPT-2-M model to generate ten documents until a text length of greater than 700 characters has been reached. This is necessary because especially the smaller GPT-2 models sometimes break off after one sentence. The upper text length is bounded by the length of the trained context of GPT-2, which is 1024 tokens including the tokens of the prompt (here: the query). After weighing up runtime and performance, GPT-2-M with 345 million parameters is always used in this thesis. The generated documents are then compared with the accumulated search results of the ChatNoir queries. This similarity factor is later used as a score for the final ranking of the results. As possible similarity metrics, the cosine similarity on the tf-idf vectors as well as on S-Bert embeddings (bert-base-nli-stsb-mean-tokens) were evaluated (see chapter 3.4), the final run used tf-idf vectors due to the limited capability of the Tira submission system.

3.3 Word based approach

Another approach for using language models to assess the relevance of documents is based on the internal understanding of models of words and their contexts. GPT-2 and other models use large transformers [17] to assign a probability to each word in their vocabulary for being the next token in the sequence. If we now condition GPT-2 on the search query, we can compare the first word of our given document with the probability predicted by GPT-2 for that word. Now we condition GPT-2 on the search query with the first word of the document as the suffix and compare the probability with the second word, and so on. Assuming an appropriate normalization over the length of the text, the sum over all word probabilities could reflect the relevance of the document, since GPT-2 can remember the meaning of the search query even over longer passages [13]. GPT-2 thus assigns low probabilities to non-topic words that it does not expect, conditioned on the search query. This is observed when GPT-2 searches are made and documents coming from a different direction are evaluated.

3.4 Evaluation

In order to evaluate whether the generated texts fit the question at all, a pairwise comparison was done between each of them. The similarity of the pairs was computed by the cosine similarity on the tf-idf vectors as well as on semantic embeddings of a language model based on BERT. We also compared the paired similarity between the generated

[1] Code can be found on Github (github.com/bjrne/gpt-question-answering)

texts between different topics. Finally, we considered the similarity of the generated texts to the first ten search results per topic.

In order to get the best texts regardless of the matching topic, different types of conditioning were tested. Based on the single question as well as on the question with pre- and suffix, texts were generated and compared qualitatively and quantitatively.

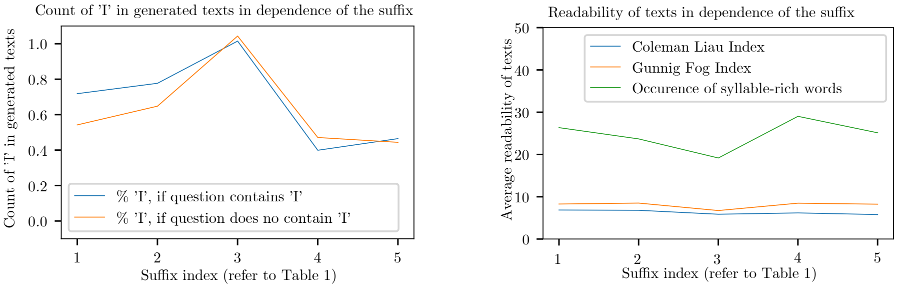
Text Generation In a manual, qualitative evaluation of text quality, it is noticeable that almost all texts deal with the right topic over the entire length. Sometimes this is at the expense of a desirable variety in the text, repetitions of the same statement are frequent. Sometimes the model ends in an infinite loop that always repeats the same phrases alternately or continuously. Furthermore, some texts do not go beyond the length of a few sentences. Sample texts can be found in Appendix 5.1.

N	Suffixes
1	<without suffix>
2	\n\n
3	\n\n I am really wondering what the answer is.
4	\n\n It is obvious that
5	\n\n Lets get to the truth about what the correct answer really is

Table 1: A list of the tested suffixes.

When comparing texts which were generated by using different suffixes, a difference in writing style becomes obvious. Texts conditioned of suffixes containing ”I” usually generate answers written in first-person perspective or personal testimonies. A quantitative analysis that corroborates this observation can be found in Figure 2a.

One can see that without adding a suffix to the query (No. 1 & 2), the perspective of the query correlates with whether the answer is given in first-person perspective or not. When using a suffix, this correlation diminishes. Moreover, it can be suspected that more objective suffixes push the generation in a less personal direction (No. 4 & 5).



(a)

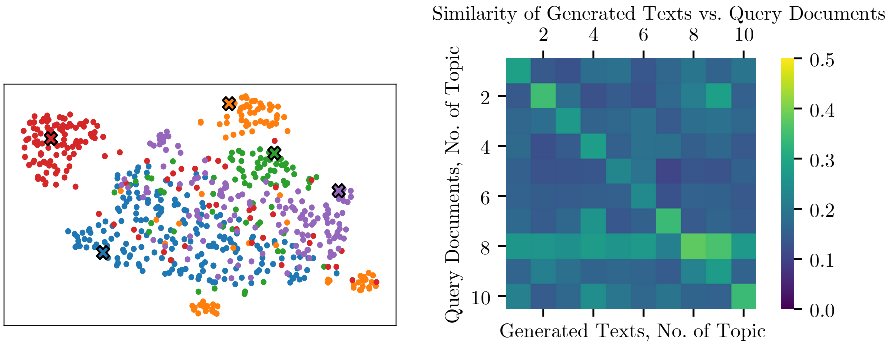
(b)

Figure 2: Left: Occurence of 'I' in GPT-2-generated texts in relation to the suffix used, see table 1. Right: Average readability in relation to the suffix used, see table 1.

Since a complete manual check of the fit of all texts to the prompt is out of scope due to both by subjectivity and effort, the generated texts were compared with each other and with search results of the same query and other topics by means of cosine similarity (see Figure 3b). In each row (generated texts) and each column (ChatNoir documents), the highest value is found in the diagonal, i.e. in the matching topic.

If you examine the generated texts using sentence embeddings generated by [14], you can see their semantic similarity to the search queries of the same topic as well as to the first documents of ChatNoir. In Figure 3a you can see the embeddings of all sentences for five topics, while the crosses mark the respective search query phrase embedding.

In a further comparison, Figure 4 shows sentence embeddings of the documents together with the texts for five topics. The dark colors represent the sentences of the generated texts, the lighter colors represent the documents. One can see in both figures that almost every sentence in the generated texts has a semantic reference to the question.



(a)

(b)

Figure 3: Left: S-Bert embeddings of all sentences of the generated documents, reduced in dimensionality by UMAP[11]. Embeddings of the queries are marked with a cross. Right: The average of the pairwise cosine-similarity of sentence-embeddings between the search result documents and the generated texts. First ten topics only.

GPT-2 Bias A major problem with this approach, which has not yet been sufficiently evaluated, is the bias present in language models. Since these are usually trained on freely available texts on the Internet, they ”learn” to adopt various implicit and explicit discriminations. If the output of such systems is now used to understand search queries, this bias is automatically passed on and ensures that the results are influenced. For example, if answers to the question ”Do I clean tiles better with a mop or vacuum cleaner” are generated and all answers and recommendations come from women, existing stereotypes are consolidated. Furthermore, new perspectives on a topic may be

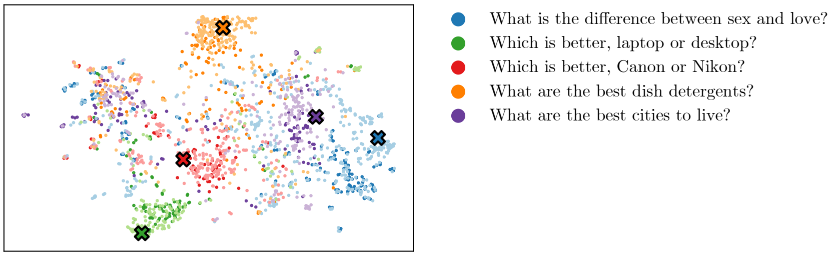


Figure 4: S-Bert embeddings of all sentences of the generated texts in light colors and search result documents in dark colors, reduced in dimensionality by UMAP[11]. Embeddings of the queries are marked with a cross.

structurally prevented if these biased texts are seen as ideal for the documents to be delivered. Other perspectives are thus likely to find it harder to be heard. A separate evaluation of the biases of this model is outside the scope of this paper, but since an unchanged, pre-trained GPT-2 model was used, relevant literature can be found in [16].

4 Conclusion

Models like GPT-2 already understand the intention of the user sufficiently well to generate suitable results. But using this knowledge to create a good ranking is not trivial. The approaches presented here show the basic possibility, but do not deliver the hoped-for results. While a semantic similarity to the research question can be recognized and thus documents with similar vocabulary can be sorted out, the ranking based on this similarity does not achieve good sorting. To recognize good arguments, other approaches seem to make more sense.

References

1. Clueweb12 - dataset. `https://lemurproject.org/clueweb12/index.php`, accessed: 6.3.2020
2. Turing-nlg - a 17 billion parameter language model by microsoft. `https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/`, accessed: 6.3.2020
3. Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Elastic chatnoir: search engine for the clueweb and the common crawl. In: European Conference on Information Retrieval. pp. 820–824. Springer (2018)
4. Bondarenko, A., Hagen, M., Potthast, M., Wachsmuth, H., Beloucif, M., Biemann, C., Panchenko, A., Stein, B.: Touché: First shared task on argument retrieval. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 517–523. Springer International Publishing, Cham (2020)

5. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. ArXiv **abs/1508.05326** (2015)
6. Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., Panchenko, A.: TARGER: Neural Argument Mining at Your Fingertips. In: Costa-jussà, M., Alfonseca, E. (eds.) 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Association for Computational Linguistics (Jul 2019), `https://www.aclweb.org/anthology/P19-3031`
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019)
8. Gomaa, W.H., Fahmy, A.: A survey of text similarity approaches (2013)
9. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B., Schmidhuber, J.: Lstm: A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems **28**, 2222–2232 (2017)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
11. McInnes, L., Healy, J.: Umap: Uniform manifold approximation and projection for dimension reduction. ArXiv **abs/1802.03426** (2018)
12. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), `http://www.aclweb.org/anthology/D14-1162`
13. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
14. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. 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). https://doi.org/10.18653/v1/d19-1410, `http://dx.doi.org/10.18653/v1/d19-1410`
15. Schildwächter, M., Bondarenko, A., Zenker, J., Hagen, M., Biemann, C., Panchenko, A.: Answering comparative questions: Better than ten-blue-links? CoRR **abs/1901.05041** (2019), `http://arxiv.org/abs/1901.05041`
16. Tan, Y.C., Celis, L.E.: Assessing social and intersectional biases in contextualized word representations (2019)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
18. 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: ArgMining@EMNLP (2017)
19. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. ArXiv **abs/1704.05426** (2017)

5 Appendix

5.1 Examples of generated text

What is the difference between love and sex? I would love to have a conversation about it.

Is love a natural and universal emotion? Or is it something that we develop and learn over time?

Is it a part of the human condition? Or is it something that we can develop and learn about through practice?

Is it something that we are born with? Or does it come through our parents?

I am curious about the nature of love and if it is a part of the human condition.

My question is a matter of life and death and that is really what it boils down to. I think all of us will be alive in our lifetimes without knowing it. We will all be alive and we will all be affected by it and have it shape us for the rest of our lives. It may not be something we understand, but we all have it.

I am wondering if love is a natural emotion, or if it is something that we develop and learn through practice?

What is the difference between sex and love?

Love, like sex, is an experience and an expression of our own desires.

Is it something that we develop and learn over time?

Yes. I think the answer is that there is a huge range of sexual experiences that can be experienced. It really depends on your level of experience and your level of experience. I think it's an experience that you are capable of.

Do we develop the same sense of love as we develop sex?

I think that sex is much more developed in your brain than love. We develop that same kind of sexual drive, the same sexual experience, that is necessary for our capacity to become human beings. That same level of experience and that same level of understanding is necessary for our capacity to have that same kind of love for one another.

Is it really necessary to be a human being, to be a loving and affectionate human being, if we are not able to have sex?

If we were born with a certain kind of sex drive, would that affect our capacity to be a good human being? If we were born with a certain kind of sexual drive, would that affect our capacity to have that same kind of love for one another?

What is the difference between sex and love?

I think sex is a natural and universal emotion.

Are there certain things that we need to learn or experience before we can know whether we are a loving and affectionate human being?

What is better, laptop or desktop? What's better? A desktop or a laptop? What's better, laptop or desktop?

I am really wondering what the answer is. Posted on Oct 18th 2017, 12:46 Reply

#21 TheLionKnight

”Killer4Life said: If I had to choose, I'd choose the desk, since I don't want to have to work with a mouse or keyboard in an environment where I could lose track of which direction to turn the mouse. The only reason I could even think of that desk is that the desk has a screen which looks really cool.

If there's no way to put a mouse on it, I might have to say no to my mouse. It's the only reason I'm choosing a desk. The only reason I could even think of that desk is that the desk has a screen which looks really cool.If there's no way to put a mouse on it, I might have to say no to my mouse. It's the only reason I'm choosing a desk. Posted on Oct 18th 2017, 13:13 Reply

22 kyren

”TheLionKnight said: If I had to choose, I'd choose the desk, since I don't want to have to work with a mouse or keyboard in an environment where I could lose track of which direction to turn the mouse. The only reason I could even think of that desk is that the desk has a screen which looks really cool.

If there's no way to put a mouse on it, I might have to say no to my mouse. It's the only reason I'm choosing a desk. Yeah, it's a good one, considering how the screen is so small, so small that you could be standing there while wearing it.

Also, yes, I'd go with the desk because it has a nice shape, the screen is awesome and the monitor is great.

I don't like how the mouse buttons are made and the size is also small.

You can always make it bigger, though.

I think the best way is to just buy the mouse, which is much cheaper, and make a stand so that you can sit on it, because it is so small.

Also, the only other thing I would change is the button layout on the side so that it isn't all vertical.

”TheLionKnight said: Yeah, it's a good one, considering how the screen is so

Which is better, Canon or Nikon? I have only used Canon lenses since I got my first camera back in 2004 and I never really cared about buying Nikon.

I am using the D800E to try the new EOS 10D (and have shot some of the best shots with it so far) so I figured I would get some opinions on the differences.

I have used both cameras and I have not found any differences in my experiences.

So I will start with the Canon EOS 30D and see if I can find any differences.

Canon EOS 30D

My first shot

Canon EOS 30D at night

Canon EOS 30D at sunrise

Canon EOS 30D at sunset

Canon EOS 30D at sunset with a flash

Canon EOS 30D at night with a flash and with a tripod

Canon EOS 30D at night with a flash and with a tripod with tripod and a flash

Canon EOS 30D at sunrise with a flash and with a tripod

Canon EOS 30D at sunset with a flash and with a tripod

Canon EOS 30D at sunrise with a flash and with a tripod and a flash

Canon EOS 30D at sunset with a flash and with a tripod and a flash

Canon EOS 30D at sunrise with a flash and with a tripod and a flash

Canon EOS 30D at sunset with a flash and with a flash

Canon EOS 30D at sunrise with a flash and with a tripod and a flash

Canon EOS 30D at sunset with a flash and with a flash

Canon EOS 30D at sunrise with a flash and with a flash

Canon EOS 30D at sunset with a flash and with a flash

Canon EOS 30D at sunrise with a flash and with a flash

Canon EOS 30D at sunset with a flash and with a flash

Canon EOS 30D at sunrise with a flash and with a flash

Canon EOS 30D at sunset with a flash and with a flash

Canon EOS 30D at sunrise with a flash and with a flash

Canon EOS 30D at sunset with a flash and with a flash

Canon EOS 30D at sunrise with a flash and with a flash

Canon EOS 30D a