LeviRANK: *Limited* Query *Expansion* with *Voting Integration* for Document Retrieval and *Ranking*

Notebook for the Touché Lab on Argument Retrieval at CLEF 2022

Ashish Rana[1,*], Pujit Golchha[1,*], Roni Juntunen[1,2,*], Andreea Coajă[1,*], Ahmed Elzamarany[1,*], Chia-Chien Hung[1] and Simone Paolo Ponzetto[1]

[1]*Data and Web Science Group, University of Mannheim, Germany*
[2]*Lappeenranta-Lahti University of Technology LUT, Finnland*

Abstract

In order to make informed decisions in personal life, the information available on the internet is often overwhelming, and thus comparative decision-making is particularly challenging. Given the plethora of online resources, it often results in wastage of resources for finding relevant and correct responses. The Touché 2022 Shared Task 2 for Comparative Questions focuses on solving this problem by retrieving corresponding documents given a comparative question. The importance of the retrieved documents is determined both by their relevance and quality. In this paper, we present a three-stage retrieval, ranking, and stance prediction system called the LeviRANK. It uses bidirectional self-attention-based language models for argumentation detection in documents. In the first stage, it incorporates an empirically novel retrieval approach that produces the highest recall values for small comparative queries. The retrieval module uses *voting*-based BM25 retrieval for merging multiple BM25 retrievals from a pool of relevant expanded queries. We then use monoT5 and duoT5 document rankers based on the *"Expando-Mono-Duo"* design pattern. Finally, we identify object stance by building a two-step stance prediction approach which first separates out documents that are specifically related to objects and further identifies the given relevant object in them. With the proposed approach, we observe that bidirectional self-attention-based document ranking models successfully identify argumentation structure better than the probabilistic ranking models. The LeviRANK system ranks the highest mean nDCG@5 score of 0.758 for document relevance task, second-highest nDCG@5 score of 0.744 for document quality task, and second-highest Macro-F1 score of 0.301 for the stance prediction task.

Keywords

Comparative Question Answering, Document Retrieval, Document Ranking, Multi-Stage Document Ranking

1. Introduction

In the current consumer-driven economic landscape with the overabundance of products and their associated information over the web, it is very hard to make informed decisions. Studies have demonstrated that for essential life decisions people prefer online research and comparative

CLEF'22: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

asrana@mail.uni-mannheim.de (A. Rana); pgolchha@mail.uni-mannheim.de (P. Golchha); roni.juntunen@student.lut.fi (R. Juntunen); acoaja@mail.uni-mannheim.de (A. Coajă); aelzamar@mail.uni-mannheim.de (A. Elzamarany); chia-chien.hung@uni-mannheim.de (C. Hung); ponzetto@uni-mannheim.de (S. P. Ponzetto)

CEUR Workshop Proceedings (CEUR-WS.org)

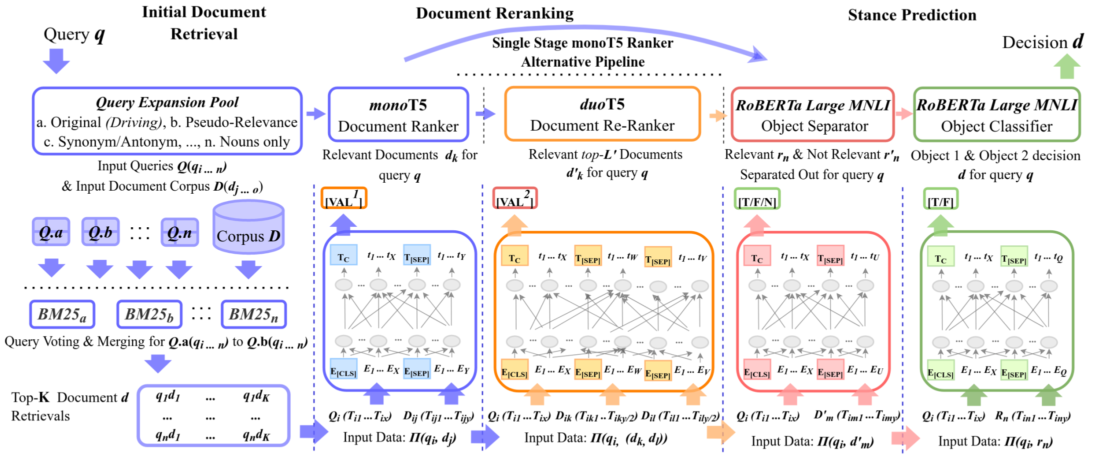

Initial Document Retrieval
Query q
Query Expansion Pool
a. Original (Driving), b. Pseudo-Relevance
c. Synonym/Antonym, ..., n. Nouns only
Input Queries Q(q_i ... n)
& Input Document Corpus D(d_j ... o)
Q.a
Q.b
Q.n
Corpus D
BM25_a
BM25_b
BM25_n
Query Voting & Merging for Q.a(q_i ... n) to Q.b(q_i ... n)
Top-K Document d Retrievals
q_1d_1 ... q_1d_K
... ...
q_nd_1 ... q_nd_K
Document Reranking
Single Stage monoT5 Ranker Alternative Pipeline
monoT5
Document Ranker
Relevant Documents d_k for query q
[VAL^1]
T_C t_1 ... t_X T_[SEP] t_1 ... t_Y
E_[CLS] E_1 ... E_X E_[SEP] E_1 ... E_Y
Q_i (T_i1 ... T_ix) D_ij (T_ij1 ... T_ijy)
Input Data: Π(q_i, d_j)
duoT5
Document Re-Ranker
Relevant top-L' Documents d'_k for query q
[VAL^2]
T_C t_1 ... t_X T_[SEP] t_1 ... t_W T_[SEP] t_1 ... t_V
E_[CLS] E_1 ... E_X E_[SEP] E_1 ... E_W E_[SEP] E_1 ... E_V
Q_i (T_i1 ... T_ix) D_ik (T_ik1 .. T_iky/2) D_il (T_il1 ... T_ily/2)
Input Data: Π(q_i, (d_k, d_l))
Stance Prediction
RoBERTa Large MNLI
Object Separator
Relevant r_n & Not Relevant r'_n Separated Out for query q
[T/F/N]
T_C t_1 ... t_X T_[SEP] t_1 ... t_U
E_[CLS] E_1 ... E_X E_[SEP] E_1 ... E_U
Q_i (T_i1 ... T_ix) D'_m (T_im1 ... T_imy)
Input Data: Π(q_i, d'_m)
RoBERTa Large MNLI
Object Classifier
Object 1 & Object 2 decision d for query q
Decision d
[T/F]
T_C t_1 ... t_X T_[SEP] t_1 ... t_Q
E_[CLS] E_1 ... E_X E_[SEP] E_1 ... E_Q
Q_i (T_i1 ... T_ix) R_n (T_in1 ... T_iny)
Input Data: Π(q_i, r_n)

Figure 1: Architectural illustration of the LeviRANK system pipeline includes three steps: initial document retrieval, relevance ranking, and stance prediction modules. The two alternative pipelines with different ranking models, namely (1) monoT5 only *(single-stage)*, and (2) monoT5-duoT5 *(multi-stage)* are highlighted in the diagram as well.

questions are also part of those decisions [1, 2]. These comparative questions can be *factual* (e.g., which footballer has the most goals?) or *contextual* (e.g., who is the best footballer?) where the retrieved *relevant* argumentative contents are often distributed over several sentences or passages [3]. The *Argument Retrieval for Comparative Questions* shared task is specially designed to handle these comparative topics, retrieve corresponding relevant documents and further give object entailment information. The distribution of context and argumentative nature of the relevant information given a question makes the document ranking and stance prediction problems especially challenging [4]. The problem becomes even more challenging when questions are purely *abstract* (e.g., what is the best way to live life?) in which the answer would vary based on personal and societal preferences. With our contribution to Touché 2022 Shared Task 2: *Argument Retrieval for Comparative Questions*, we intend to explore these questions and detect argumentative structures in relevant documents with a self-attention mechanism [5].

For designing our initial document retrieval we use the well-expanded passage corpus with queries generated by DocT5Query provided on the TIRA forum [6, 7, 8]. Based on our analysis of the recall metric, it is observed that query drift is quite prevalent when we are using *ad-hoc* versions of different supplementary retrieval approaches (e.g., pseudo-relevance, query expansion) leading to lower recall values. Hence in this work, we propose **L**imited Query **E**xpansion with **V**oting **I**ntegration for Document Retrieval and **RANK**ing (LeviRANK), a novel framework of argument retrieval for comparative questions. The LeviRANK system, depicted in Figure 1 includes three steps, namely (1) *Initial Retrieval:* this consists of an empirically designed query expansion variant that starts with an initial BM25 retrieval from a limited pool of expanded queries [9, 10, 11]. This fixed pool of expanded queries, is prepared by limiting the different query expansion methods to either *replace* (e.g., adjectives with their synonym/antonym pairs), *add* (e.g., pseudo-relevance queries), or *remove* (e.g., noun-only queries) only one term to restrict query drift. Finally, we use the original query's BM25 retrieval documents as a driving relevance set for top-1000 relevant documents. We further utilize voting amongst a pool of queries to select the most relevant documents from this 1000 document set to prepare a more concise document relevance set. After that, for each remaining query in the query pool, we append top-retrieved disjoint document sets to our earlier prepared concise document relevance set in a cascading manner. And the corresponding set size for the top-retrieved documents depends on the relevance of each query which is manually tuned for this task by us. (2) *Document Ranking:* we next utilize the "Expando-Mono-Duo" design pattern for two-stage pointwise and pairwise document ranking from T5 language models [6, 12]. (3) *Stance Prediction:* we use two-step classification RoBERTa language models [13] to handle the unbalanced multi-class stance prediction problem. In the first step we segregate out documents containing object-relevant information and in the second step, we identify the specific object's relevance in that document given a topic query.

[1] The contribution of our team *Captain Levi* to this paper is to provide an empirically relevant retrieval approach with limited query expansion for comparative queries. We also show that our demonstrated retrieval approach is more representative of the relevant document space for a given topic query. Additionally, we also investigate the representational capabilities of the bi-directional self-attention-based monoT5 and duoT5 document ranking models. Finally, we also quantify the stance prediction capabilities of the two-step multi-class classification approach both in zero-shot and fine-tuned settings for object entailment tasks.

2. Related Work

Argument mining, document retrieval and ranking tasks have been extensively studied with successful deep learning approaches in recent years [4, 14, 15, 16, 17]. Previously, the *Argument Retrieval for Comparative Questions* task was focused on retrieving relevant argumentative passages from generic web crawl document collections [18, 19]. In related experimental studies, the approaches have primarily utilized ChatNoir by inputting either the original or expanded preprocessed queries for initial document retrieval [20]. And post the initial retrieval, the documents have been ranked by using multiple machine learning and deep learning approaches like random forest, XGBoost, LightGBM, Word2vec, GPT-2, fine-tuned BERT and DistilBERT [21, 22, 23, 24, 25, 26, 27, 28, 29]. In our approach, we additionally tackle the initial retrieval problem by using the text passages expanded with queries generated using DocT5Query instead of using ChatNoir [7, 6].

Industry and academia multi-stage ranking retrieval systems have arguably been one of the most practical solutions for modern search systems [30, 31, 32]. Multi-stage retrieval and document ranking pipelines with dynamic embedding representations obtained from bi-directional language model architectures have achieved great results in the past [4, 14, 15]. The bi-directional self-attention architecture in BERT successfully attends to important tokens and captures the semantic relationship between them [33]. Additionally, upgrading to dynamic masking where different sentence components are masked per epoch increases the robustness and performance as shown by the language models like RoBERTa [34]. Additionally, Text-To-

[1] All resources developed as part of this work are publicly available at: https://github.com/softgitron/LeviRank.

Text Transfer Transformer (T5) model with a large pre-trained dataset called the Colossal Clean Crawled Corpus (C4) also gives state-of-the-art results. It, unlike BERT-style language models, takes text data both as inputs and outputs. Additionally, with respect to unsupervised pre-training stage innovations like causal prediction task, deshuffled original input text prediction and masked tokens prediction also gives T5 additional performance boost [35].

In recent studies, monoBERT- and duoBERT-based pointwise and pairwise document relevance models have achieved promising results on MS MARCO dataset [36, 37]. Additionally, the design pattern has been generalized and dubbed as *"Expando-Mono-Duo"* with T5 re-ranking models for pointwise and pairwise comparison at consecutive stages [6]. Here, "Expando" refers to document passage expansion with generated queries by DocT5Query model trained on MS MARCO passage ranking task. The "Mono" and "Duo" indicate the pointwise and pairwise comparisons for document ranking. In this work, we investigate the performance and utility of each component with *"Expando-Mono-Duo"* pattern. The stated ranking models are not specifically pre-trained on comparative argument retrieval questions but rather on general-purpose MS MARCO ranking dataset queries. This leaves these models biased towards making good ranking predictions for certain topic queries, but not for all the topic queries under consideration for this task. We develop our document retrieval and ranking system by empirically improving upon the limits of this design pattern specific to this use case.

Stance prediction formulated as an entailment problem has helped in analyzing different problems like political discourse, scientific misinformation, and comparative questions understanding [14, 15, 38]. Additionally, bi-directional self-attention-based models like BERT, RoBERTa and T5 have shown promising results in both regular and zero-shot learning settings [15, 13, 39]. Unbalanced stance prediction problems can be broken into two-stage multi-class classification problems in order to improve distinguishing capabilities amongst classes for smaller datasets [13]. For the LeviRANK framework, we investigate this two-step classification approach by first separating object classes with {No, Neutral, Object} as stance prediction objects and then further classify separated objects into {First, Second} classes. Here, the {No, Neutral, Object} labels highlight the absence, neutrality and presence of the object-related information in the relevant documents corresponding to the given topic query. Additionally, the {First, Second} labels highlight if any object in the relevant document is the answer to the given comparative topic query. For the topic queries present in the *Argument Retrieval for Comparative Questions* shared task, zero-shot learning performance is measured with Macro-F1 metric. But, we also additionally use a 50/50 test/train split on worst and best query topics respectively from the zero-shot learning task to further fine-tune our models and analyze the performance improvement with the Macro-F1 metric.

3. Datasets

The dataset of the Touché 2022 Shared Task 2 consists of: (1) a collection of 868,655 passage documents extracted from the ClueWeb12 where the average document length is approximately 150 words and (2) 50 queries on different topics [40]. These documents can either be *non-relevant*, *relevant*, or *highly relevant* for any given topic query. For stance prediction, an additional dataset with 956 comparative questions and answers [38] are provided. It consists of object detection and classification labels generated from annotating data subsets from Stack Exchange and Yahoo Answers topic classification datasets.

In this task, we are given a topic query set $\mathcal{Q}$ and a relevant document corpus $\mathcal{D}$. For each query $q \in \mathcal{Q}$, our aim is to retrieve all uniquely relevant documents $d \in \mathcal{D}$ and help categorize them as *y(q,d)* in {First, Second, No, Neutral} for object stance prediction. Here, *y(q,d)* represents the oracle stance prediction function given a query q and document d as inputs. The performance of the retrieval subtask is determined by *nDCG@5* (i.e., Normalized Discounted Cumulative Gain over top-5 highest scored documents) [41, 42]. And for the stance prediction problem, the performance is evaluated by *Macro-F1* score metric which it derives an averaged-out metric assigning each class an equal weight.

4. Methodology

We formulate and divide subtasks for the *Argument Retrieval for Comparative Questions* task into three stages, namely: *retrieval*, *ranking*, and *stance prediction.* Figure 1 depicts the complete proposed LeviRANK pipeline and the following subsections explain the proposed system in detail. To make informed decisions about the individual implementation components for each subtask we devise our own alternative evaluation strategy which is elaborated in the below subsection.

4.1. Initial Evaluation Strategy for Subtask Module Selection

We formulate and divide our evaluation metrics into three stages of the system, namely: *retrieval*, *ranking*, and *stance prediction.* For the retrieval stage, we focus on increasing the Recall@K metric value which measures the maximum number of relevant documents retrieved out of all the possible relevant documents for K number of total retrieved documents by a given system. For the ranking stage, we use the nDCG@5 metric which provides us quantitative insights into what fraction of the top-5 relevant documents gets retrieved in the correct order for the given topic queries. Finally, for the stance prediction evaluation, we consider Macro-F1 score for the entailment task architecture where Macro-F1 weighs F1-score obtained for each class equally.

We combine the gold standard labels from the previous two Shared Task 2 iterations having annotated document relevance ranking information. The LeviRANK system's retrieval and ranking stage performance gets evaluated over 100 topic queries from the past two years. Additionally, to match the document identifiers exactly from the gold standard, we use all the unique ChatNoir urls from this year's corpus to scrape data and build a new corpus where average document length gets increased to is approximately 4500 words. Although for the previous task iterations the document size is quite large, we believe evaluation of our system on this very large document setup gives us a lower bound on the LeviRANK system's ranking performance. Because, for the ranking T5 language architectures maximum input document token length is 512 & 256 tokens respectively for monoT5 & duoT5 respectively. Therefore, large web documents of approximately 4500 words get truncated before the majority of argumentation structures can be attended over by self-attention. Finally, for evaluating the stance prediction model we directly use the stance prediction dataset and compare our Macro-F1 results with the existing baseline solution [38].

Table 1
Initial document retrieval metric performance for comparative query topics with Recall@K metric where K={1000, 1500, 2000} represents number of retrieved documents.

Retrieval Approach	Recall@1000	Recall@1500	Recall@2000
BM25 Baseline	**90.18**	90.67	91.11
Dense Retrieval	85.70	86.56	87.56
Pseudo-relevance Feedback	89.98	90.59	91.07
LeviRank Voting	90.14	**91.08**	**91.17**

4.2. Initial Retrieval

Our retrieval problem mathematically summarizes to gathering of relevant documents from corpus $\{d_j, ..., d_n\} \in \mathcal{D}$ for topic queries $q \in \mathcal{Q}$. First, we do basic preprocessing *(e.g., lowercasing after punctuation, query tags and NLTK based stopword removal, WordNet based lemmatization)* for the documents and then build the probabilistic BM25 index for these documents with Pyserini package [43, 44, 9]. The corpus $\mathcal{D}$ used for building index and ranking documents is already expanded with queries generated using DocT5Query. In order to improve Recall@K we additionally try out query expansion and relevance feedback approaches [45]. Continuously, we build a custom dense-representation-based index that utilizes ColBERT representations which were initially fine-tuned for MS MARCO dataset [46]. By comparing average Recall@K scores for each retrieval approach, we quantify the retrieval limitations of each of the approaches listed above. Additionally, with individual query level analysis we observe that different retrieval approaches perform better for different topic queries and the Recall@K trend is not consistent for the topic queries for these retrieval approaches.

For this task, we observe that adding, replacing, or removing a few words causes a large query drift. The phenomenon is attributed to the small length of the query often containing a maximum of two nouns and one adjective as a comparator. Hence, with our proposed approach we generate nine new queries and fetch their corresponding BM25 retrieval results. These nine queries are expanded in a limited manner that only adds, removes, and replaces one specific word from the original query. Also, we limit the influence of all these queries by only adding a proportion of their individual BM25 retrieval document list based on their corresponding relevance determined by us experimentally.

In the LeviRANK system, the original query is the main driving query from which maximum documents are retrieved and added to the final retrieval document set. To remove irrelevant retrieval results from the initial 1000 documents, we further ensure that each of these documents has at least one existing retrieved copy in the other retrievals from the expanded query pool. This is done by iteratively checking complete retrieval sets from all the expanded queries for each document belonging to the top-1000 document set of the original query. Essentially, meaning that we take input votes for every document in the initial 1000 documents from the query pool retrieval sets to obtain the most relevant set of documents.

The following document sets are then appended in a cascading manner: (1) Two disjoint document sets of 150 most relevant retrieval documents each from two queries, first which contain only nouns and another one from which only stopwords are removed; (2) A disjoint document set of 60 most relevant documents each from 3 queries in which synonym replaces the comparative adjective clause; (3) A disjoint document set of 30 most relevant documents each from 2 most relevant antonyms queries with the replaced adjective. Until the 2000 document retrieval count is reached, (4) the pseudo-relevance feedback retrieval document set that is not part of the current extended retrieval document set is appended in equal proportion with the original query's remaining retrieval set disjointly. With the union of such a diverse set of disjoint document retrievals in their respective proportions based on manually assigned query relevance, we intend to make our retrieval results set close to (i.e., more representative) the ideal retrieval results. Table 1 demonstrates the better performance of our approach with higher values of Recall@1500 and Recall@2000.

4.3. Document Ranking

In this subtask, relevant documents $\widehat{\mathcal{D}}_K(q,d)$ = { $d_1(q,d), ..., d_K(q,d)$ } are ranked, where each document comprises of sentences { $s_1(q,d), ..., s_Y(q,d)$ }. Here, $\widehat{\mathcal{D}}_K(q,d)$ represents documents from the initial retrieval. Also, $s_1(q,d)$ *and* $s_Y(q,d)$ denote the index of sentences from *1 to Y*. We use all the top-K *(K=2000)* retrieved documents for each query from the previous subtask to rank them with monoT5's default implementation from PyGaggle library without fine-tuning [47]. The given language model receives input sequence $<q_i$,*[SEP]*,$d_j>$ and produces binary true/false label target tokens for the documents. At inference time, for probability computations the model outputs a condensed single relevance score using softmax of the true/false label logits for each document. The monoT5 performs pointwise comparison between the documents. Additionally, we use duoT5 model to further granularize the document ranking with pairwise comparison.

For the second re-ranking stage we use the top-k *(k=100)* relevant documents from the mono-T5 stage and again rank them with the duoT5 model. In this re-ranking subtask, relevant documents $\widehat{\mathcal{D}}_k(q,d)$ = {$d_1(q,d), ..., d_k(q,d)$} are ranked, where each document comprises of sentences { $s_1(q,d), ..., s_{Y/2}(q,d)$ }. Here, $\widehat{\mathcal{D}}_k(q,d)$ represents documents after the monoT5 ranking stage. $s_1(q,d)$ *and* $s_{y/2}(q,d)$ are the index of sentences from *1 to Y/2* which is half of original document length of 512 tokens. The given language model receives input sequence as $<q_i$, *[SEP]*,d_k, *[SEP]*,$d_l>$ and outputs are given by single relevance score using SYN-SUM method given by the equation 1. Here, in the equation 1, i represents the given document and s_i equals the document's given score. Also, j gives a document for comparison and $\mathcal{J}_i$ represents document comparisons for document i. Finally, $p_{i,j}$ and $p_{j,i}$ represents pairwise score for document i compared with document j and vice-versa respectively.

$$\text{SYS-SUM: } s_i = \sum_{j \in J_i} (p_{i,j} + (1 - p_{j,i})) \tag{1}$$

As the results shown in Table 2, monoT5 model performs the best for the document ranking task on previous year topic queries with the given merged documents corpus. We also observe that the average nDCG@5 values for BM25, monoT5-only, monoT5-duoT5 are 0.33, 0.47, and 0.31 respectively. With our individual query level analysis based on nDCG@5 metric scores, duoT5 model performance is inconsistent while handling large documents. The input sequence

Table 2
Document ranking results of different ranking approaches with their corresponding nDCG@5 scores.

Ranking Approach	BM25	monoT5-only	monoT5-duoT5
nDCG@5	0.33	**0.47**	0.31

Table 3
Stance prediction F1-score results of the LeviRANK system compared with the most accurate sentiment prompt-based system result from Bondarenko et al. [38].

Approach	No object	Neutral	Object 1	Object 2	Macro-F1
Bondarenko et al. [38]	0.40	**0.53**	0.70	0.63	0.57
LeviRank	0.40	0.52	**0.72**	**0.68**	**0.58**

size of the duoT5 model reduces to 256 tokens for both document sequences under analysis in comparison to the 512 tokens of monoT5 model, which results in loss of relevant argumentation information. Additionally, with our manual analysis, it is indicated that the starting section of the web documents from the corpus often contains irrelevant information (e.g., links, headers), which further contributes to the performance drop of the duoT5 model for this subtask.

4.4. Stance Prediction

In this subtask, we use the predicted relevant documents $\widehat{\mathcal{D}'}_k(q,d) = \{d\ '_1(q,d), \ldots, d\ '_m(q,d)\}$ from the ranking retrieval stage to predict the object stance $\hat{y}(q,d)$ of the comparative queries $q \in \mathcal{Q}$. Here, $\widehat{\mathcal{D}'}_k(q,d)$ represents documents ranked after the second reranking stage. Additionally, $\hat{y}(q,d)$ and $y(q,d)$ represent the predicting and oracle stance functions of the given document-query pairs. We formulate this subtask as a two-stage binary classifier problem where the first classifier separates and predicts the documents with $\hat{y}(q,d)$= {No, Neutral} and {Object} labels with input sequence $<q_i,[SEP],d\ '_m>$. And, the second classifier predicts the stance $\hat{y}(q,d)$={First, Second} with input representation $<q_i,[SEP],r_n>$. The pre-trained RoBERTA-Large-MNLI language model is used for *Object Separator* predictions and pre-trained RoBERTA-Large-MNLI is chosen for predicting the final *Object Stance*.

The results in Table 3 demonstrate the advantage of using the two-step binary classification process for the LeviRANK system given unbalanced small datasets in comparison to traditional four-way multi-class classification. We attribute this performance gain to better prediction of the Object labels with high F1-score and Recall scores of 84.4% and 93.93% respectively. Multi-class classification models perform poorly for predicting undersampled classes due to their label scarcity in the dataset. Random oversampling with replacement of the underrepresented classes combined with cross-validation is further implemented for increasing prediction capabilities amongst No and Neutral classes. Hence, by using these techniques the LeviRANK's two-step

Table 4
Submission result summary from leaderboard of the Touché Shared Task 2: Argument Retrieval for Comparative Questions for the LeviRANK system.

Submitted Approaches	Recall@2K	Input$_{\text{duoT5}}$	nDCG@5$_{\text{rel}}$	nDCG@5$_{\text{qual}}$
TCT-ColBERT+monoT5+duoT5	92.05	100	**0.758**$_1$	**0.744**$_3$
BM25+monoT5+duoT5	98.23	100	0.755$_2$	0.742$_4$
LeviRANK+PR+monoT5+duoT5	97.96	50	0.753$_3$	0.730$_5$
LeviRANK+monoT5	**98.34**	0	0.727$_4$	0.706$_6$
Pseudo-Relevance*(PR)*+monoT5	97.16	0	0.722$_5$	0.695$_7$

classification approach reduces false positive predictions of No and Neutral classes against the Object class. Additionally, our proposed approach performs significantly better for First and Second object labels which makes it even a better alternative. Since, for comparative topic queries it is highly important to know which object is being discussed in either positive or negative light within the very highly relevant documents.

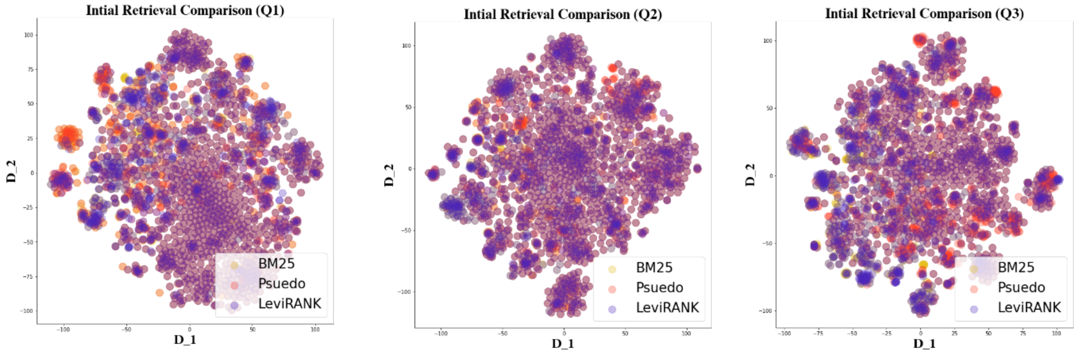
5. Results and Error Analysis

The leaderboard results for the Touché Shared Task 2: Argument Retrieval for Comparative Questions is summarized on the Table 4. Our submission set included five systems in total, where three of them use the monoT5-duoT5 ranking & reranking architecture whereas the other two only include monoT5 ranking architecture. The monoT5 only architecture was selected as a fallback architecture approach because of unexpected duoT5 ranking results for large documents. The specific module component details for each submission and respective subtask is highlighted in the *Submitted Approaches* column of Table 4. From Table 4, we observe that the LeviRANK initial retrieval approach achieves the highest Recall@2000 score value as expected. But, we also infer that the input size to duoT5 model is one of the major result driving factors for obtaining the highest nDCG@5 score. Additionally, by comparing TCT-ColBERT [48] and BM25 based systems we can observe that the higher Recall@2000 score doesn't necessarily guarantee best nDCG@5 results for both quality & relevance metric score. Also, quality and relevance requires separate model architectural design for performance improvement. Even though our models achieve the highest nDCG@5 scores for relevance, they still miss out on producing documents of the highest quality.

For stance prediction subtask our approach obtained the Macro-F1 score of 0.301 with the two-step classification approach. This sub-par performance can be especially attributed to the low performance while predicting No and Neutral classes. Since the evaluation dataset was not available during the stance prediction model, we can interpret this result as zero-shot learning performance obtained after training on the Stack Exchange & Yahoo Answers topic classification datasets. Further, for measuring the stance prediction capabilities of our two-step stance prediction approach we fine-tune our entailment language models on the annotations corresponding to the top-50 % best query topics. Where top-50 % best query topics represents

Table 5
Stance prediction Macro-F1 score results of the LeviRANK system measuring (1.) Zero-shot performance on the complete annotated relevant documents dataset, (2.) Zero-shot performance on the annotations corresponding to top-50 % worst performing topics, (3.) Performance evaluation on the annotations corresponding to top-50 % worst performing topics after fine-tuning on the annotations corresponding to the top-50 % best performing topics.

Training Approach	Prediction Annotations Set	Macro-F1
Zero-shot Two-Step RoBERTA-MNLI architecture	Whole stance dataset	0.303_2
Zero-shot Two-Step RoBERTA-MNLI architecture	Worst 50 % topic queries	0.116_{6^*}
Two-Step RoBERTA-MNLI *(fine-tuned, 50 % best topics)*	Worst 50 % topic queries	$\mathbf{0.387_{1^*}}$



Q1. What is better Google search or Yahoo search?

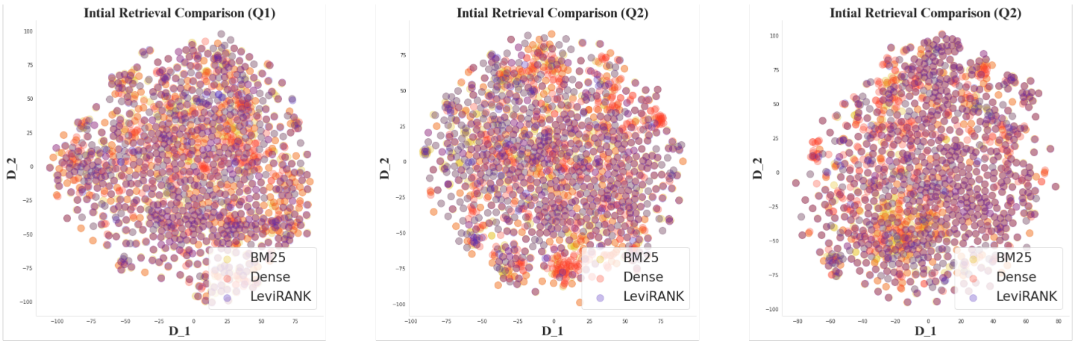
Q2. Which is better MAC or PC?

Q3. Which one is better Netflix or Blockbuster?

Figure 2: Qualitative latent representation comparison between retrieved *large size* documents for topic queries at *the Initial retrieval* stage, demonstrating more representationally spreaded initial retrieval of the LeviRANK system.

the top-25 topics for which our entailment models achieved the best Macro-F1 score. And, obtain the highest Macro-F1 score of 0.387 for the unseen annotations on query topics for which our system's performance was reportedly the worst, as shown in Table 5. This improvement in performance can directly be attributed to addition of new training annotations in the RoBERTa-MNLI language models.

As illustrated in Figures 2 and 3 with the t-SNE (t-distributed Stochastic Neighbor Embedding) representation plots, it is encouraging to see that the 2000 document set representational spread from our proposed system the LeviRANK is far higher as compared to the baseline BM25 and relevance-feedback-based retrieval results [49]. Also, the t-SNE plots are concerned with pairwise distances between the points and attempts to visualize high-dimensional data in a low-dimensional 2D space. Here, the individual axes in the t-SNE have no quantifiable meaning and these plots are referred for qualitative analysis. Additionally, it is observed that the higher latent vector space spread is valid for query topics belonging to multiple different topic domains as well. We argue that our proposed methodology successfully retrieves documents that contain a high degree of variances amongst themselves irrespective of the document size alongside

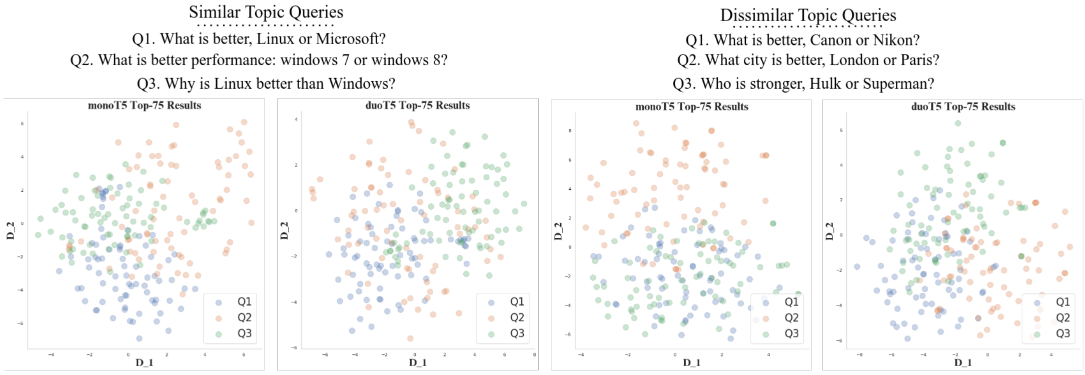


Q1. What is better Google search or Yahoo search?

Q2. Which is better MAC or PC?

Q3. Which is better Family Guy or The Simpsons?

Figure 3: Qualitative latent representation comparison between retrieved *regular size* documents for topic queries at *the Initial retrieval* stage, again demonstrating more representationally spreaded initial retrieval of the LeviRANK system.

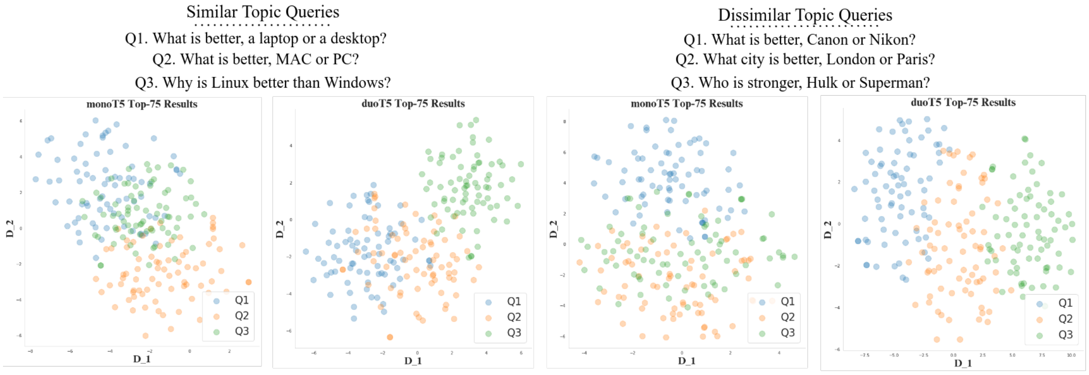


a. monoT5 Ranker on similar queries. b. duoT5 Ranker on similar queries. c. monoT5 Ranker on dissimilar queries. d. duoT5 Ranker on dissimilar queries.

Figure 4: Latent representation comparison between ranked documents by the monoT5 and duoT5 models at the *large document ranking* stage demonstrates strong document distinguishing capabilities amongst top-ranked documents for different queries.

capturing most relevant documents. This gives the retrieval performance boost when a very large number of documents are retrieved for our proposed multiple retrieval-based voting and merging approach as evident by highest Recall@2000 scores.

When considering experiments with large documents, inconsistent behavior is observed for the duoT5 model where for some queries it performs substantially well but its ranking capabilities suffers in general leading lower nDCG@5 score than monoT5. But, for regular size document sequences as shown in Table 4 this issue doesn't reproduce itself. As depicted in Figures 4 and 5, we analyze the ranking behavior of monoT5 and duoT5 on groups of similar and dissimilar queries with respect to the top-75 document set retrievals. From the ranking t-SNE plots of both the models on similar and dissimilar queries, it is clear to see that top-ranked documents have a more separated clustered structure for monoT5 model as compared to duoT5 in Figure 4 and vice-versa for Figure 5. We further argue that for large documents, superior distinguishing capabilities amongst the top-ranked documents by monoT5 model, in general,



a. monoT5 Ranker on similar queries. b. duoT5 Ranker on similar queries. c. monoT5 Ranker on dissimilar queries. d. duoT5 Ranker on dissimilar queries.

Figure 5: Latent representation comparison between ranked documents by the monoT5 and duoT5 models at the *regular document ranking* stage demonstrates stronger document distinguishing capabilities amongst top-ranked documents for different queries.

are the reason for these relatively disjoint clusters and their better performance. Specifically, since the retrieval document corpus of most relevant documents is the same for both models, duoT5 is biased towards selecting particular sets of large documents leading to a reduced ability to produce disjoint clusters for different topic queries. This performance analysis is strictly limited to this large document scenario. Since we find the opposite results in Figure 5 where the duoT5 model for both similar and dissimilar queries produces more distinct clusters proving its superior distinguishing capabilities.

6. Conclusion and Future Work

In this work, we propose the LeviRANK system, which uses multi-stage reranking architecture to rank relevant documents for comparative questions. For this system, we implement a novel retrieval approach that systematically merges retrieval results from the restricted query pool based on voting. Additionally, retrieval results are appended in a cascading manner where the appended retrieval result size depends on the relevance assigned to the query. This retrieval approach also attempts to find synergy amongst multiple retrieval techniques like relevance feedback, query expansion, and docT5query for improving Recall@2000 result values. This cascading retrieval merging approach achieves the highest Recall@2000 values of 91.17 and 98.42 for the combined previous two years and current Touché Shared Task 2's comparative topic queries respectively.

We further investigate the performance of the "Expand-Mono-Duo" design pattern for the *ad hoc* retrievals. For the Touché Task 2: Argument Retrieval for Comparative Questions, our ranking pipeline obtains the best performance of 0.758 for nDCG@5 metric in the relevance evaluation task and the second-best performance of 0.744 for nDCG@5 metric in the quality evaluation task. With these results, we conclude that bi-directional self-attention models successfully capture comparative argumentative structure for given topic queries especially for medium-length documents in a pairwise document comparison setting. Further, we observe that our system suffers in ranking performance when document size becomes large, especially the duoT5 model. This performance decrease is attributed to a lack of argumentation structure being present within the maximum input token length limit for these T5 language model architectures. For the stance prediction task, our model achieves a Macro-F1 score of 0.301 which is lower than the Macro-F1 obtained in the dev-set of the stance prediction dataset. This decrease in performance can be attributed to the especially low prediction performance on the No and Neutral labels.

The LeviRANK system in summary provides the best relevant document results out of all the existing systems. And further, gives competitive performance while predicting the stance of the retrieved documents. For future work, we intend to systematically study the causes of inconsistencies in document relevance & quality assessment ranking results amongst different query topics and qualitatively produce more accurate & consistent ranking systems.

Acknowledgments

We would like to extend our gratitude towards our professor Dr. Simone Paolo Ponzetto and tutor Chia-Chien Hung for sharing their valuable experience and insights with us. We also appreciate the contribution efforts of Touché team for providing all the extended support and help.

References

[1] E. Turner, L. Rainie, Most americans rely on their own research to make big decisions, and that often means online searches (2020).

[2] A. Bondarenko, P. Braslavski, M. Völske, R. Aly, M. Fröbe, A. Panchenko, C. Biemann, B. Stein, M. Hagen, Comparative web search questions, in: Proceedings of the 13th International Conference on Web Search and Data Mining, 2020, pp. 52–60.

[3] H. Trivedi, H. Kwon, T. Khot, A. Sabharwal, N. Balasubramanian, Repurposing entailment for multi-hop question answering tasks, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2948–2958. URL: https://aclanthology.org/N19-1302. doi:10.18653/v1/N19-1302.

[4] J. Lawrence, C. Reed, Argument mining: A survey, Computational Linguistics 45 (2020) 765–818.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[6] R. Pradeep, R. Nogueira, J. Lin, The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models, 2021. URL: https://arxiv.org/abs/2101.05667. doi:10.48550/ARXIV.2101.05667.

[7] R. Nogueira, J. Lin, A. Epistemic, From doc2query to docttttquery, Online preprint 6 (2019).

[8] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1_5.

[9] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, R. Nogueira, Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations, in: Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021), 2021, pp. 2356–2362.

[10] K. S. Jones, S. Walker, S. E. Robertson, A probabilistic model of information retrieval: development and comparative experiments: Part 2, Information processing & management 36 (2000) 809–840.

[11] S. Robertson, H. Zaragoza, et al., The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, arXiv preprint arXiv:1910.10683 (2019).

[13] X. Zeng, A. Zubiaga, Qmul-sds at sciver: Step-by-step binary classification for scientific claim verification, arXiv preprint arXiv:2104.11572 (2021).

[14] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074. doi:10.18653/v1/N18-1074.

[15] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7534–7550. URL: https://aclanthology.org/2020.emnlp-main.609. doi:10.18653/v1/2020.emnlp-main.609.

[16] S. MacAvaney, A. Yates, A. Cohan, N. Goharian, Cedr: Contextualized embeddings for document ranking, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 1101–1104.

[17] R. Nogueira, Z. Jiang, J. Lin, Document ranking with a pretrained sequence-to-sequence model, arXiv preprint arXiv:2003.06713 (2020).

[18] A. Bondarenko, M. Fröbe, M. Beloucif, L. Gienapp, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2020: Argument Retrieval, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 11th International Conference of the CLEF Association (CLEF 2020), volume 12260 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2020, pp. 384–395. URL: https://link.springer.com/chapter/10.1007/978-3-030-58219-7_26. doi:10.1007/978-3-030-58219-7_26.

[19] A. Bondarenko, L. Gienapp, M. Fröbe, M. Beloucif, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2021: Argument

Retrieval, in: K. Candan, B. Ionescu, L. Goeuriot, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association (CLEF 2021), volume 12880 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2021, pp. 450–467. URL: https://link.springer.com/chapter/10.1007/978-3-030-85251-1_28. doi:10.1007/978-3-030-85251-1_28.

[20] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Elastic chatnoir: Search engine for the clueweb and the common crawl, in: European Conference on Information Retrieval, Springer, 2018, pp. 820–824.

[21] V. Chekalina, A. Panchenko, Retrieving comparative arguments using ensemble methods and neural information retrieval, Working Notes of CLEF (2021).

[22] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[23] T. Abye, T. Sager, A. J. Triebel, An open-domain web search engine for answering comparative questions., in: CLEF (Working Notes), 2020.

[24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems 30 (2017).

[25] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[27] T. K. H. Luu, J.-N. Weder, Argument retrieval for comparative questions based on independent features (2021).

[28] A. Alhamzeh, M. Bouhaouel, E. Egyed-Zsigmond, J. Mitrović, Distilbert-based argumentation retrieval for answering comparative questions, Working Notes of CLEF (2021).

[29] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[30] N. Asadi, J. Lin, Effectiveness/efficiency tradeoffs for candidate generation in multi-stage retrieval architectures, in: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013, pp. 997–1000.

[31] R.-C. Chen, L. Gallagher, R. Blanco, J. S. Culpepper, Efficient cost-aware cascade ranking in multi-stage retrieval, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 445–454.

[32] S. Liu, F. Xiao, W. Ou, L. Si, Cascade ranking for operational e-commerce search, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1557–1565.

[33] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[35] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu,

et al., Exploring the limits of transfer learning with a unified text-to-text transformer., J. Mach. Learn. Res. 21 (2020) 1–67.

[36] R. Nogueira, W. Yang, K. Cho, J. Lin, Multi-stage document ranking with bert, arXiv preprint arXiv:1910.14424 (2019).

[37] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, Ms marco: A human generated machine reading comprehension dataset, in: CoCo@ NIPS, 2016.

[38] A. Bondarenko, Y. Ajjour, V. Dittmar, N. Homann, P. Braslavski, M. Hagen, Towards Understanding and Answering Comparative Questions, in: K. S. Candan, H. Liu, L. Akoglu, X. L. Dong, J. Tang (Eds.), 15th ACM International Conference on Web Search and Data Mining (WSDM 2022), ACM, 2022, pp. 66–74. URL: https://dl.acm.org/doi/10.1145/3488560.3498534. doi:10.1145/3488560.3498534.

[39] R. Pradeep, X. Ma, R. Nogueira, J. Lin, Scientific claim verification with VerT5erini, in: Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, online, 2021, pp. 94–103. URL: https://aclanthology.org/2021.louhi-1.11.

[40] A. Bondarenko, M. Fröbe, J. Kiesel, S. Syed, T. Gurcke, M. Beloucif, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2022: Argument Retrieval, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2022, p. to appear.

[41] W. B. Croft, D. Metzler, T. Strohman, Search engines: Information retrieval in practice, volume 520, Addison-Wesley Reading, 2010.

[42] C. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval, Natural Language Engineering 16 (2010) 100–103.

[43] S. Bird, Nltk: the natural language toolkit, in: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, 2006, pp. 69–72.

[44] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.

[45] W. B. Croft, S. Cronen-Townsend, V. Lavrenko, Relevance feedback and personalization: A language modeling perspective., in: DELOS, Citeseer, 2001.

[46] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 2020, pp. 39–48.

[47] R. Pradeep, R. Nogueira, J. Lin, Pygaggle, 2021. URL: https://github.com/castorini/pygaggle.

[48] S.-C. Lin, J.-H. Yang, J. Lin, Distilling dense representations for ranking using tightly-coupled teachers, arXiv preprint arXiv:2010.11386 (2020).

[49] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (2008).