Overview of the Causality-driven Adhoc Information Retrieval (CAIR) task at FIRE-2020

Suchana Datta
University College Dublin
suchana.datta@ucdconnect.ie

Debasis Ganguly
IBM Research Europe, Dublin, Ireland
debasis.ganguly1@ie.ibm.com

Dwaipayan Roy
IISER, Kolkata, India
dwaipayan.roy@gmail.com

Derek Greene
University College Dublin
derek.greene@ucd.ie

Charles Jochim
IBM Research Europe, Dublin, Ireland
charlesj@ie.ibm.com

Francesca Bonin
IBM Research Europe, Dublin, Ireland
fbonin@ie.ibm.com

ABSTRACT

This paper describes an overview of the findings of the track named 'Causality-driven Ad hoc Information Retrieval' (abbv. CAIR) at the Forum for Information Retrieval Evaluation (FIRE) 2020. The purpose of the track was to investigate how effectively can search systems retrieve documents that are causally related to a specified query event. Different from standard information retrieval (IR), the criteria of relevance in this search scenario is stricter in the sense that the retrieved documents at the top ranks should provide information on the potentially relevant causes that might have caused a given query event, e.g. retrieve documents on political situations that might have led to 'Brexit'. We released a dataset comprised of a set of 25 queries split into train and test sets. We received submissions from two participating groups. The two main observations from the best performing runs from the two participating groups are that longer queries showed a general trend to yield more causally relevant documents towards top ranks as seen from the results obtained from the first participating group, whereas it turned out that sequence-based text representation for semantically matching the documents with queries did not yield effective retrieval results, thus leaving the scope to develop supervised or semi-supervised methods to address causality-based retrieval.

ACM Reference Format:
Suchana Datta, Debasis Ganguly, Dwaipayan Roy, Derek Greene, Charles Jochim, and Francesca Bonin. 2020. Overview of the Causality-driven Adhoc Information Retrieval (CAIR) task at FIRE-2020. In *Forum for Information Retrieval Evaluation (FIRE '20), December 16–20, 2020, Hyderabad, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3441501.3441513

1 INTRODUCTION

In traditional ad hoc IR setup, a search system retrieves a ranked list of documents given a query. The usefulness of the output of an ad hoc IR system, in the form of a ranked list of documents, is limited in situations when i) decision makers need to formulate policies to mitigate a current event that requires attention (e.g. drop in the value of British pound), or ii) policy-making regarding societal benefits (e.g. formulating government policies to reduce housing crisis by analyzing the main likely causes). In the aforementioned situations, a traditional search system user is required to carefully analyze the topically relevant documents (likely to describe the main event expressed in the query itself) and most likely needs to reformulate queries in order to retrieve documents related to the potential *causes leading* to the (query) event.

As an example, if a user would like to find potential causes leading to the 'drop of British pound' (and the user is not aware of these causes, i.e. the search intention is to explore rather than recalling previously known information), he first needs to enter a query related to the event itself (an example query could be 'pound value drop'). The documents retrieved at top ranks by a traditional search system will mostly be on this topic itself (since these documents are expected to have high term weight values for the query words), e.g. recent news reporting the drop in the value of the pound. Since such top ranked documents retrieved by a traditional IR model are not likely to be *causally relevant* (listing the likely causes leading to the query event) to the information need, the user then needs to manually reformulate his queries by including terms that are representative of the likely causes (e.g. concepts such as 'Brexit delay', 'negotiation difficulties between EU and UK' etc.).

The user of a traditional IR system, hence, needs to spend considerable effort in reformulating queries in order to retrieve the causally relevant documents towards top ranks. In this track, we seek to investigate approaches to reduce this manual effort and ask participants to design effective retrieval models seeking to address *causality-based relevance* rather than the traditional *topical relevance.*

2 CAIR TASK

Motivated by the scenario described in Section 1, we proposed a shared task in the FIRE 2020 track. We provided participants a static test collection of 303291 news documents and a list of 25 queries, divided into two parts - 5 queries for training and 20 queries for test, related to events that were likely to be caused by a number of other past events. We also provided associated relevance judgements for the set of train queries. The participants were then required to develop ranking models that could effectively retrieve documents containing information on such past events which were likely candidates to lead to the query event. Proposed model from each participating team must generate a 6 column .tsv file following

FIRE '20, December 16–20, 2020, Hyderabad, India

ACM ISBN 978-1-4503-8978-5/20/12…$15.00
https://doi.org/10.1145/3441501.3441513

the standard TREC format. In order to encourage the investigation of different kind of features, we allowed three runs per participating group. The officially submitted ranked lists of different participating systems were then evaluated by comparing them against a set of manually judged relevant documents.

3 DATASET

As the notion of causality differs from the idea of topical relevance, the selection of topics for this task was restricted to the query events with causal information need. To illustrate the dataset characteristics of CAIR task, we present the difference in relevance (topical (R_T) and causal (R_C)) in Table 1 for a selected topic *'Assassination of Osama-bin-laden'* from CAIR dataset.

3.1 Target Collection

We chose a static test collection of news articles constituting the official English ad hoc IR collection of FIRE [8] as our target collection. The news articles were crawled from the source 'Telegraph India'[1] published over a period of 10 years (2001 to 2011). The crawled content is formatted with XML markup into separate categories or domains, such as 'sports', 'business', 'international' etc. The total number of documents in the collection is over $300k$ (303291 to be precise).

3.2 Query Formulation

Considering the fact that our target collection was a newspaper corpus having news over the time span $2001-2011$, we searched for different significant events occurred during the same period which might have a series of leading causes. But having chosen topics this way, we faced difficulties while retrieving documents containing causality. As we intended to capture chain of causes of an event, mere word matching did not work. So then we started studying the topic from different sources, like - online newspapers, blogs, editorial articles etc. and tried to gather potential keywords indicating causes of the particular event. We further used those keywords to reformulate queries which eventually led us to a potential chain of causes for the deemed query event. We also made use of a subset of the FIRE adhoc query set hosted by [2] and eventually we compiled 25 queries in total having causal information need.

Each topic follows the standard TREC format, i.e., is comprised of a *title* (usually a small number of keywords), a *description* (a well-formed sentence describing the information need in more details) and a *narrative* (a paragraph describing the relevance criteria in details). While selecting the topics, we took the following into consideration.

(1) We ensured that a query is representative of an event that occurred during the period covered by the target collection, i.e. between 2001-2011.
(2) An event qualifies as a valid topic only if there exists a multiple number of potential (arguable) causes that might have led to it. We eliminate those cases where the notion of causality is mentioned in the same document also describing the query event or it does not help user to walk through the chain of query event at all.

[1] https://www.telegraphindia.com/

3.3 Relevance Assessments

In practice, we combine a number of top-retrieved document set obtained from different IR models with distinct setup to create the pool of relevance judgements [9]. However, for the CAIR task, this traditional technique of making pool was very unlikely to be proven useful. There were two main reasons behind this, such as -

(1) As the notion of causality is quite different in nature from that of topical notion, it is quite evident that we can not blindly depend on traditional IR models, such as - LM, BM25 etc. Also, unlike topical IR, we do not have any existing aid that is empirically proved to be applicable for causal IR. It is worth mentioning that, one of the notable purpose of building the judgement pool is to mitigate this gap in causality-driven IR research.
(2) In contrast to traditional IR (topical), causality-based judgement requires assessors to be well-versed with the event given in the query, i.e. assessors must have some initial knowledge on the query event so as to identify the connection amongst series of cause-effect events.

Therefore, to construct relevance assessments for our target 25 queries, we used our existing knowledge about the news events to manually collate a number of (causally) related articles. In particular, a standard search system (Lucene with different retrieval and pseudo-feedback model configurations) was used to submit a series of manually reformulated queries to construct a pool of documents to judge. We made use of the outcome of LM, BM25 and Relevance-based LM to do a series of query reformulations in order to capture causal trails.

The reformulations were based on a combination of existing knowledge of the assessors and the content of the top-ranked documents for the submitted query. Assessors then checked each document in the pool and assigned a binary score (0 being non-relevant and 1 being relevant) depending on their assessment of the causal link between the information present in the document (e.g. 'Brexit uncertainty') with that of the query ('pound fall'). In addition to this, we manually judged outcomes received from participating teams to further extend the pool of relevance assessment. We judged only the distinct subset of the documents (absent in the initial pool) retrieved by each reported model.

3.4 Train-Test Splits

Note that the CAIR task can be addressed primarily in two ways, (a) either as an unsupervised retrieval task in which for a given query event, user's search intention is to capture only causally relevant documents from the whole collection, or (b) it can be modelled as a classification task where the objective is to discriminate causally connected documents from that of topical set, considering the fact that causal set of documents are likely to be a subset of the topical set of retrieved documents.

Therefore, initially we released 5 topics with relevance judgements as *training data* and furthermore, we provided 20 other topics as *test set*. The total set of relevance pool for all 25 queries is also available at [1] for the participants which might help them in further self-assessment.

Query - Assassination of Osama-bin-Laden	
Topical	Pakistan's President Asif Ali Zardari today said that the whereabouts of Al Qaida leader Osama bin Laden remained a mystery...
RelDoc: 1	was a suspicion that he could be dead... Zardari said US officials had told him that they had no trace of the Al Qaida chief.
	...a leaked foreign intelligence document published....a loud buzz that Osama bin Laden may have died of typhoid in Pakistan last month, but no country would confirm anything...
RelDoc: 2	...citing an uncorroborated report from the Saudi secret services that the leader of al Qaida terror network had died. The chief of al Qaida was a victim of a severe typhoid crisis while in Pakistan on August 23, 2006, the document said...
Causal	An audio tape broadcast... sounds like the voice of Osama bin Laden threatening attacks against US allies,... If it genuinely is bin Laden's voice, makes references to recent events such as last months Bali bombings and the Chechen hostage siege in Moscow...
RelDoc: 1	warned US allies that they would be targets of new attacks... The United States blames bin Laden and his Al Qaida network for the September 11, 2001, hijacked plane attacks on America that killed more than 3,000 people, ...
	Osama bin Ladens al Qaida network may be plotting spectacular attacks inside the US,... Bin Laden and Al Qaida have been blamed by Washington for the hijacked aircraft attacks on September 11, 2001, which killed about 3,000 people...
RelDoc: 2	Al Qaida may favour spectacular attacks that meet several criteria: high symbolic value, mass casualties, severe damage to the US economy and maximum psychological trauma, the FBI said...

Table 1: Excerpts of relevant documents (both topical and causal) for a query seeking information on Obama's assassination.

4 MODELS PROPOSED BY PARTICIPATING TEAMS

We received a total of *four* submissions from *two* participating teams this year. The model architecture of each group is described below:

- **UCSC** [3]: This team was from University of California, Santa Cruz. The participating team proposed a causal connection detection model with the help of query expansion technique. The team claimed that their proposed model is a simplified version of the work presented in [10] and goes through several steps, such as -
  - *Event Extraction* : They first extracted events from the title of the query.
  - *First Retrieval* : With the help of extracted events, the team performed an initial retrieval with a hope to capture potential pre-events (prior events related to the query event).
  - *Causal Relation Detection* : They used causal keywords (e.g. 'because', 'after', 'lead to' etc.) to observe if two sentences are causally connected.
  - *Query Expansion* : Top 5 candidate pre-events are chosen for expanding the initial query.
  - *Second Retrieval* : The second retrieval is performed with the expanded query and they report it as final ranked retrieved result set.

  UCSC team submitted a total of 3 runs. They reported *post-event-terms-expansion* as their proposed model, whereas *query-narratives* and *query-title* have been reported as potential baselines. However, query-narratives method outperformed all other reported models as depicted in Table 2.
- **NITS** [7]: Team 'NITS' was from National Institute of Technology, Silchar, India. They used sequence-based Universal Sentence Encoder (USE) [4] for word embedding where the model is trained by a deep averaging network encoder. News article documents were first split into small chunks to encode and then those chunks were represented in a word embedding space. Cosine similarity was used for capturing similarity between a query and document retrieved. However, result shows that sequence-based text representation for semantically matching the documents with queries did not yield effective retrieval results.

5 EVALUATION

Each participating team was allowed to submit at most three runs. Team UCSC has submitted three runs while one run was submitted by team NITS. We evaluate each submitted run based on their performance achieved over 20 test queries. In particular, we used the following evaluating measures to report model's efficiency:

- **MAP:** We chose Mean Average Precision (MAP) as our primary measure of retrieval effectiveness so as to take both precision and recall into account. This metric quantifies the retrieval model based on the mean of the average precision scores achieved per query.
- **P@5:** We also made use of $P@5$ to measure model's efficacy, i.e. number of relevant documents present in the top 5 ranked documents and averaged over test query set.

The performances are evaluated using 'trec-eval' [2] and the results are reported in Table 2.

6 RESULTS

Table 2 describes the performance of participating teams at a glance. It is observed that longer queries showed a general trend to yield more causally relevant documents towards top ranks as reported by the UCSC team as baseline (i.e. *query-narratives*) which emphasizes the fact that short queries are less likely to identify the trail of causality. As claimed by authors in [5], causally connected documents are likely to have only a partial term overlap with the corresponding topical set, query narrations are certainly a good resource of finding such causality specific terms given the query event. Team NITS proposed a supervised approach with USE, however the model dis not achieve good MAP which eventually indicates that supervised or semi-supervised approaches are worth exploring.

[2] https://trec.nist.gov/trec_eval/

Team Name	Run ID	MAP	P@5	Model Summary
UCSC	**query-narratives**	**0.4553**	**0.7000**	detect causal relations, query expansion
	query-title	0.4066	0.5400	
	post-event -terms-expansion	0.3885	0.5000	
NITS	run-1	0.0577	0.2600	embedding with USE [4]

Table 2: Retrieval effectiveness of models proposed by participating teams (best performing model outcome is boldfaced).

7 CONCLUDING REMARKS

The CAIR task comprises both the immediate goals of the initial run from the participants and longer term goals from better understanding of causal search as the task evolves. The goals of the first iteration was to establish a common understanding of causal search and a common platform for evaluating that. Other domains of study that look at causal events match events in very narrow contexts (e.g., both events must occur in a headline [6]). It is important to study causality using IR techniques so that we can properly explore causality over whole documents and more importantly across documents [5]. We believe that one of the products of this task is lists of queries and relevant documents that defines the connections between documents containing potential causes to a query event. Also, this common platform allows us to determine which approaches work for causal search and which do not, and it also allows us to confirm that this area of search is sufficiently different from ad hoc IR as to warrant study.

Although we do not expect an IR system to definitively prove causation, these systems could provide output that allows us to reason on the causation or simply correlation of different events. This type of output could be used in reasoning systems or aid in constructing structured knowledge sources around how events are connected (e.g., *causally* in addition to temporally or topically). Causal search, as we are framing it, fits well in the common search paradigm used IR, and it will require those techniques more common to IR to find connections across documents and not connect only those events found in a single document or narrow span of text.

REFERENCES

[1] 2020. Causality-driven Adhoc Information Retrieval. https://cair-miners.github.io/CAIR-2020-website/.

[2] 2020. Forum for Information Retrieval Evaluation. http://fire.irsi.res.in/fire/2020/home.

[3] Lin C. and Zhang Y. 2020. Causality Detection for Causality-driven Adhoc Information Retrieval Task. In *Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation (December 2020).*

[4] 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., and Kurzweil R. 2018. Universal Sentence Encoder. arXiv:cs.CL/1803.11175

[5] S. Datta, D. Ganguly, D. Roy, F. Bonin, C. Jochim, and M. Mitra. 2020. Retrieving Potential Causes from a Query Event. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, 1689–1692. https://doi.org/10.1145/3397271.3401207

[6] C. Hashimoto, K. Torisawa, J. Kloetzer, Motoki Sano, I. Varga, J. H. Oh, and Y. Kidawara. 2014. Toward Future Scenario Generation: Extracting Event Causality Exploiting Semantic Relation, Context, and Association Features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 987–997. https://doi.org/10.3115/v1/P14-1093

[7] Dadure P., Pakray P., and Bandyopadhyay S. 2020. Preliminary Investigation on Causality Information Retrieval. In *Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation (December 2020).*

[8] Palchowdhury S., Majumder P., Pal D., Bandyopadhyay A., and Mitra M. 2011. Overview of FIRE 2011. In *Multilingual Information Access in South Asian Languages - Second International Workshop, FIRE 2010, Gandhinagar, India, February 19-21, 2010 and Third International Workshop, FIRE 2011, Bombay, India, December 2-4, 2011, Revised Selected Papers*. 1–12.

[9] E. Voorhees and D. Harman. 2000. Overview of TREC-8. In *Proc. of TREC-8*. 1–23.

[10] S. Zhao, Q. Wang, S. Massung, B. Qin, T. Liu, B. Wang, and C. Zhai. 2017. Constructing and Embedding Abstract Event Causality Networks from Text Snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Cambridge, United Kingdom) *(WSDM '17)*. Association for Computing Machinery, New York, NY, USA, 335–344. https://doi.org/10.1145/3018661.3018707