CAIR - Causality-driven Adhoc Information Retrieval

1 Introduction & Motivation

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. Taking this into consideration, 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 Why we need such System?

In contrast to a traditional search system, a causal search system (CSS) seeks to retrieve documents that provide information on the likely causes leading to a query event. In this extended search system, in addition to the topically relevant ranked list of documents, the user will also be presented with a list of

causally relevant documents. On submitted queries pertaining to an event (e.g. 'drop of pound' or 'housing crisis'), the system then retrieves adequate information required to construct further analysis for the purpose of automated (or semi-automated with humans-in-loop) decision and policy making. Moreover, information extracted from causally related documents could also serve as the necessary *explanations* in order to support an automatically generated decision prescribing ways to eradicate a likely cause.

3 Causal Retrieval Dataset Characteristics

A dataset for the standard IR ad hoc retrieval task is comprised of three components, namely a) a document collection, b) a set of queries, and c) relevance assessments for each query.

In relation to the first component, i.e., the document collection, it can be reasoned that the task of causal retrieval is mainly appealing for a collection of news documents, which typically describe different points-of-views on contemporary events, such as elections, economy, sports etc. Expert views and analysis of a number of these contemporary events typically forecast likely directions in which the current state-of-affairs could lead to. Consequently, it is likely that news articles from the past could contain information that describe the likely causes leading to a present event.

The queries for the causal retrieval task should correspond to those which specifically describe an event in time, e.g. 'outbreak of a war between two or more nations', 'major economic crisis' etc. Events for which there is a single cause, which is rather evident in nature (e.g., the cause is revealed in the article about the effect itself), are not interesting from the perspective of the causal retrieval task definition. Some concrete examples of such a direct cause-effect relationships are: i) news about massive rainfall in a region accompanied with the news about flooding in certain localities; ii) news about mass shooting by a gunman followed by the news on his arrest, etc. In contrast to these direct cause-event relationships, we in this work are rather interested in those cases where pieces of causal relations are spread across a number of different articles with multiple opinions on subject matters open to different interpretations, e.g., it is difficult to find a single direct cause for the drop in the pound value (prior to Brexit).

The criteria for the relevance assessments is different for the causal retrieval case, since the relevance of a document in this case is judged by whether the information in the document relates to a potential cause of the effect specified in the given query. Fig.1 illustrates the differences between the two types of relevance for a sample query seeking information on the assassination of Osama bin Laden. While it is seen that the notion of *traditional relevance* corresponds to the topic itself (the two sample topical relevant documents describe the possibility of bin Laden's death), the sample *causally relevant* documents contain information about a number of events that eventually might have been responsible for bin

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.
RelDoc: 2	...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... ...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, ...
RelDoc: 2	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... 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...

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

Laden's death (e.g. 'Bali bombings', 'hijacked aircraft attacks which killed more than 3000 people', 'severe damage to US economy' etc.).