Causal Document Retrieval: An Analytical Investigation

Pankaj Dadure, Partha Pakray and Sivaji Bandyopadhyay

Department of Computer Science & Engineering, National Institute of Technology Silchar, Assam, India

Abstract

In recent times, the extraction of semantic relation has become extremely useful for the tasks of information retrieval, question answering, decision making, and event prediction. There are a number of relationships such as cause-effect, if-then, part-whole, and etc., that express essential information about how different events or entities are anticipated in relation to one another. Cause-effect relationships, in particular, are considered to play an important role in human cognition due to their ability to leverage decision-making. In this paper, we have investigated the potential of neural network-based language representation models such as the BERT model and the Apache Nutch for the task of causal documents retrieval. The BERT model transformed the sequence of words into fixed-size embedding vector and used cosine similarity for relevance measurement. In comparison, the Apache Nutch is just a keyword matching approach and used an AND search module to retrieve news articles that match the user's query. When it came to performance, the BERT model was the worst, while Apache Nutch showed some positive results.

Keywords

Information Retrieval, Causation Relation, BERT, Apache Nutch

1. Introduction

Information Retrieval (IR) [1] is one of the well-researched fields of Natural Language Processing (NLP) which aims to retrieve the relevant information from a huge amount of data. The general IR system takes the user's query as an input, works on the similarity estimation, and returns the rank of relevant search results. This is a common principle used by today's retrieval systems. Most of the retrieval systems considered only the keywords matching instead of considering the semantic meaning of the words and sentences. It has been seen that the mapping of the semantic relations between the words and sentence yield better retrieval results [2]. In many applications, causation relation is the most considerable semantic relation [3][4]. Researcher from the NLP community has used the hand-coded pattern to derived causation relation from natural text [5]. There are two components to any causal construction: the cause and its effect. For example, "People cut down trees", as s result, "Fewer baby birds are hatched". In this example, the cause is represented by "People cut down trees", and the effect by "fewer baby birds are hatched". Cause-and-effect constructions in English can be either explicit or implicit. The explicit causal

Forum for Information Retrieval Evaluation, December 13-17, 2021, India

krdadure@gmail.com (P. Dadure); parthapakray@gmail.com (P. Pakray); sivaji.ju.cse@gmail.com (S. Bandyopadhyay)

0000-0002-8003-9671 (P. Dadure); 0000-0003-3834-5154 (P. Pakray)

patterns include relevant keywords like “cause”, “effect”, “consequence”, “induce”, and so on. The implicit causal patterns rely on semantic analysis and background knowledge.

There are many different causes of a problem that can be investigated using the cause-and-effect (CE) diagrams of Ishikawa [6]. A CE diagram can be used as a guide for allocating the resources and making the necessary investments to fix the problem. Despite the fact that, there is a lack of analytical methodologies to support these diagrams construction. A relevance feedback model [7] aims to depict the distribution of causal terms that are fairly uncommon but associated with significant weights in the topical relevance model. These terms are highly related to the topically relevant distribution, which gives rise to a variability from topical relevance towards potential causal relevance. Pundit algorithm is an well recognized predictor for future events caused by the present events [8]. To attains this, the system has been trained with examples of causality relations that have been extracted from the Internet. The trained collection contains the 150-year-old news reports. The pairs of structured events that are supposed to be related by causality are identified using textual causality patterns (such as "X because Y", "X causes Y", and so on), where the outcome is a semantically structured causality graph with 300 million fact nodes and more than one billion edges. Afterwards, the designed pundit algorithm has uses the large ontologies to generalised the causality pairs and predict causality of unseen cases.

In the CAIR-2020 task [9], we have investigated the efficacy of the universal sentence encoder model to retrieve the documents which holds cause-effect relation [10]. In the CAIR-2021 task, we have analyzed the behaviour of the bidirectional encoder representations from transformers model over the simple keyword matching approach. This paper aims to explored the potential of a pre-trained language representation model (such as the BERT model) to retrieve causality-related documents. In addition to this, we have also explored the simple AND search mechanism of Apache Nutch software. The experimental results indicate that the pre-trained language representation model needs refinement to handle the semantic structure of the longer-size text. The paper is structured as follows: Section 2 gives a detailed account of the dataset. Section 3 provides detailed description about the system architecture. Section 4 describes the experimental results. Section 5 concludes with summary and directions of further research and developments.

2. Dataset

The dataset contains the 3,03,291 news articles which are extracted from the Telegraph India[1]. The metadata about the dataset is given in table 1. The dataset contains two fields namely doc_id, and text where text is the main body or content of the news articles. The process of generating a causally related dataset is time and effort-consuming. The queries in a causality-based information retrieval system hold a cause-effect relationship with needed information.

3. System Architecture

In the proposed methodology, we have used two different frameworks to extracts casual relations between documents. The first framework that we used is the BERT-based embedding model

[1] https://cair-miners.github.io/CAIR-2020-website/

Table 1
Corpus Description

Source	News Articles from Telegraph India
Formats	Text
Size	1.69 GB
No. of Articles	3,03,291
No. of Test Queries	20

[11][12] and the second one is Apache Nutch[2]. At the initial stage, all the sequences of words contained in the news articles have been transformed into the lower case. This is because, when we transformed the words into vector form, the words with different case representations hold the different vector representations. For example, the phrases "Babri Masjid demolition case against Advani" and "babri masjid demolition case against advani", both the phrases have the same semantic and syntactic meaning with different case representations. However, the multi-dimensional vector space model considers them as two different phrases.

3.1. BERT

The preprocessed news articles (context of news article in lower case) have been fed into the BERT model and creates the vector representation (embeddings) for the same. The BERT (Bidirectional Encoder Representations from Transformers) [11] is implemented on the Transformer architecture to learn the contextual relationships between words. A transformer is an encoder-decoder-based architecture that transforms one sequence to another. When generating an embedding-based representation, BERT only needs the encoder part to perform the encoding. The sequence of tokens is fed into the BERT encoder, which transforms them into vectors and then processes in the neural network model. In general, the BERT model transformed the sequence of words into fixed sized embedding vector. In order to begin with processing, BERT's data must first be altered and embellished with additional metadata which is pictorially represented in 1:

- **Token embeddings:** Each sentence begins with a [CLS] token and ends with a [SEP] token.
- **Segment embeddings:** Each token is marked with an indicator indicating Sentence A or Sentence B. In this way, the encoder is able to discern the difference between sentences.
- **Positional embeddings:** In order to indicate the token's position in the sentence, each token has a positional embedding.

After a successful embedding of the sequence of words contained in the news articles and user entered query, we have used cosine similarity measure [13][14] to estimate the similarity between the news articles and user entered query. As a final search result, the top-20 news articles have been retrieved based on the higher similarity score.

[2]http://nutch.apache.org/

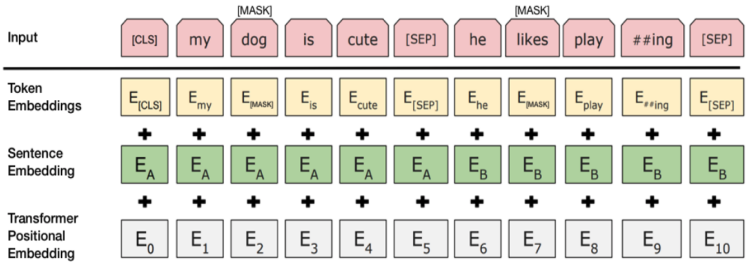

Figure 1: BERT input representation

3.2. Apache Nutch

In addition to the BERT-based embedding model, we have also used Apache Nutch [15] to retrieve the causality-related news articles. Apache Nutch is a highly extensible and scalable open-source web crawler software. It uses the Apache Lucene as its indexing core, Nutch crawler, and Nutch query search module equipped with distinguished AND search features. The workflow of the Apache Nutch is shown in figure 2. In this section, we have described their system architecture and step-wise algorithmic working principle:

1. In the first step, the name of the news articles are iteratively read into a text file (urls.txt). Each of these articles' names associates with a news article contained in the corpus.
2. In the second step, the injector module collects the document's name from the urls.txt file and uses them as seed URLs. Afterward, the CrawlDb entry is created by injecting seed URLs into CrawlDb. Prior to crawling, the injector module checks URLs and confirms the valid regex patterns defined in the nutch configuration file (regex-urlfilter.txt). We can modify the regex patterns to enable or disallow URLs designated to a specific file type using an injector.
3. As a result of CrawlDb, the generator creates a list of URLs entitled Fetchlist and stores it into the segment directory.
4. Fetcher acquired the contents of the URLs from fetchlist and stored them back into the segment directory.
5. Afterward, the parser module parsed the fetched contents and saved them back to the segment directory. As a result, a segment directory consists of fetchlist, fetched contents, and parsed contents.
6. Subsequently, the segment directory content update by the updater.
7. Afterward, links are inverted to give preference to inlinks (the number of pages pointing to the current document) over outlinks (number of pages to which the document is pointing) by using the Link Inverter module. As a result of LinkDb's storage of inlink information, ranking can be improved.
8. For each segment in the segment directory, repeat steps 3–7.
9. CrawlDb, LinkDb, and Segment Directory information are used by Apache Lucene Indexer to create an index.
10. An index has been used by the basic AND search module to retrieve a set of news articles that match with the user's query.

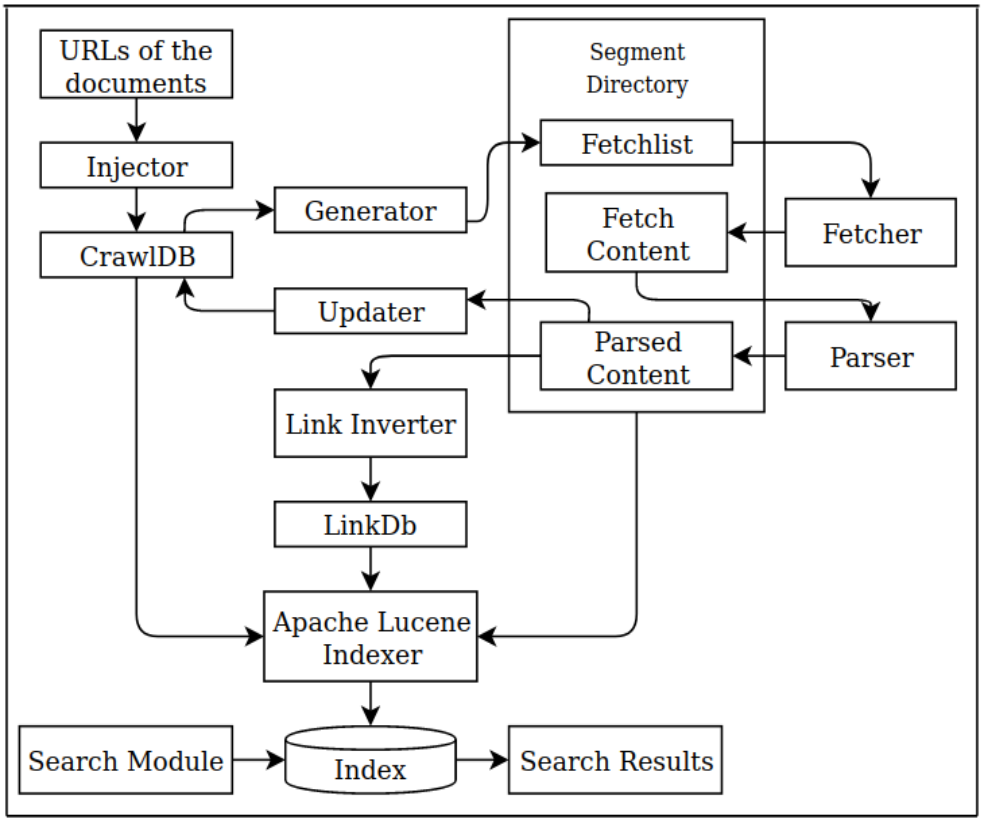

Figure 2: Workflow of the Apache Nutch

4. Experimental Results

To evaluate the performance of these approaches, we have used news articles of the Telegraph India. The obtained results value for the both approaches are shown in table 2. The results value indicates that the designed BERT model needs improvement to understands the casual relation between documents. Moreover, the Apache Nutch architecture performed quite better as compared to language representation model (such as BERT). The proposed methods have been evaluated using 20 queries, each of which consists 'title' (usually a small number of keywords) and a 'narrative' (a paragraph describing the relevance criteria in detail). For each query, we have retrieves the top 20 articles. To depict the proficiency of the proposed approaches, MAP is the primary measure [16]. The effectiveness of the proposed approaches is also calculated in terms of P@5 for more generalized and comparative analysis [17]. The obtained results for the queries "Assassination of Osama-bin-laden" and "Court blocks facebook in pakistan" are shown in table 3 and 4. The obtained retrieved results have highlighted the following points:

- The BERT embedding model is not able to provide the coherent semantic representations to long text sequence, worse fits than existing embedding approaches, and has failed to

inferred the meaningful semantic relationships between long text sequence.

- In the proposed BERT-based embedding approach, we have applied the BERT model to the entire document and generates the single embedding vector. Consequently, the obtained results values are negligible.
- The straightforward working principle of the Apache Nutch performed fairly good for the task causal relation extraction.
- In causal relation extraction, sometimes the terms that indicating the cause are not present in the query. In that case, the keyword matching approach is not sufficient to retrieve the causality-related documents.

Table 2
Experimental results

Approach	MAP	P@5
BERT		
Query based on title	0.0014	0.0100
Query based on narrative	0.0001	0.000
Apache Lucene	0.1063	0.4800

Table 3
Retrieved Results for Query "Assassination of Osama-bin-laden"

Sr. No.	Snippet of the retrieved articles	Article Id
1	The al Qaida network has claimed responsibility for attacks on an Israeli airliner and hotel in Kenya which killed 16 people and vowed even more lethal assaults against Israel and its chief ally, the US.	1021209_foreign_story_1462594.utf8
2	Pakistans most feared Islamic militant group, branded by Washington last week a foreign terrorist group, has been severely weakened by a crackdown on extremism, intelligence officials said today.	1030208_foreign_story_1651225.utf8
3	Crown Prince Abdullah said Saudi Arabia would triumph over evil powers in its war against terror after warnings by the US and Britain of new terror threats in the kingdom. The de facto ruler of Saudi Arabia, birthplace of al Qaida leader Osama bin Laden, warned people not to back terrorists.	1030815_foreign_story_2267415.utf8

5. Conclusions and Future Scope

In this paper, we have studied two different notions to retrieves the documents, which hold semantic relations such as cause-effect. In cause-effect relation, there is no direct relation between the information and the topic of a query but involves the information that could have led to the query event. To identify this relation, we have used pre-trained BERT language

Table 4
Retrieved Results for Query "Court blocks facebook in pakistan"

Sr. No.	Snippet of the retrieved articles	Article Id
1	A Pakistani court on Wednesday ordered the telecommunication authorities to block Facebook for inviting its members to post caricatures considered blasphemous.	1100520_foreign_story_12468810.utf8
2	Pakistani authorities today blocked video sharing website YouTube for hosting objectionable content, a day after cutting off access to Facebook over a page featuring blasphemous caricatures of Prophet Mohammed.	1100520_frontpage_story_12469628.utf8
3	A prominent Iranian cleric has appeared on state radio to declare that the fatwa issued against Salman Rushdie in 1989 is still alive. Ahmad Khatami made his provocative comments during Friday prayers amid heightened protests across the Muslim world against the authors knighthood.	1070623_foreign_story_7962343.utf8

representation model and Apache Nutch. The obtained results value has indicated that the BERT language representational model is insufficient to map the cause-effect relationship between the documents. Moreover, the Apache Nutch is just a keyword matching model, but it still performed better than the BERT model. In the future, we will work with a more robust language representational model such as Sentence-BERT to embedded the long textual data more precisely and concisely.

Acknowledgment

The authors would like to express their gratitude to the Department of Computer Science & Engineering and Center for Natural Language Processing, National Institute of Technology Silchar, India for providing the infrastructural facilities and support.

References

[1] A. Singhal, et al., Modern information retrieval: A brief overview, IEEE Data Eng. Bull. 24 (2001) 35–43.
[2] Z. Xu, X. Luo, S. Zhang, X. Wei, L. Mei, C. Hu, Mining temporal explicit and implicit semantic relations between entities using web search engines, Future Generation Computer Systems 37 (2014) 468–477.
[3] P. Menzies, Is causation a genuine relation?, in: Real Metaphysics, Routledge, 2003, pp. 128–144.
[4] S. Datta, D. Greene, D. Ganguly, D. Roy, M. Mitra, Where's the why? in search of chains of causes for query events., in: Proceedings of The 28th Irish Conference on Artificial

Intelligence and Cognitive Science, Dublin, Republic of Ireland, December 7-8, volume 2771 of *CEUR Workshop Proceedings*, 2020, pp. 109–120.

[5] N. Asghar, Automatic extraction of causal relations from natural language texts: a comprehensive survey, arXiv preprint arXiv:1605.07895 (2016).

[6] R. U. Bilsel, D. K. Lin, Ishikawa cause and effect diagrams using capture recapture techniques, Quality Technology & Quantitative Management 9 (2012) 137–152.

[7] S. Datta, D. Ganguly, D. Roy, F. Bonin, C. Jochim, M. Mitra, Retrieving potential causes from a query event, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1689–1692.

[8] K. Radinsky, S. Davidovich, S. Markovitch, Learning causality for news events prediction, in: Proceedings of the 21st international conference on World Wide Web, 2012, pp. 909–918.

[9] S. Datta, D. Ganguly, D. Roy, D. Greene, C. Jochim, F. Bonin, Overview of the causality-driven adhoc information retrieval (cair) task at fire-2020, in: Forum for Information Retrieval Evaluation, 2020, pp. 14–17.

[10] P. Dadure, P. Pakray, S. Bandyopadhyay, Preliminary investigation on causality information retrieval, in: FIRE (Working Notes), 2020, pp. 771–779.

[11] 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).

[12] A. Yates, R. Nogueira, J. Lin, Pretrained transformers for text ranking: Bert and beyond, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 1154–1156.

[13] P. Dadure, P. Pakray, S. Bandyopadhyay, An empirical analysis on retrieval of math information from the scientific documents, in: International Conference on Communication and Intelligent Systems, Springer, 2019, pp. 301–308.

[14] F. Rahutomo, T. Kitasuka, M. Aritsugi, Semantic cosine similarity, in: The 7th International Student Conference on Advanced Science and Technology ICAST, volume 4, 2012, p. 1.

[15] R. Khare, D. Cutting, K. Sitaker, A. Rifkin, Nutch: A flexible and scalable open-source web search engine, Oregon State University 1 (2004) 32–32.

[16] E. Yilmaz, J. A. Aslam, Estimating average precision with incomplete and imperfect judgments, in: Proceedings of the 15th ACM international conference on Information and knowledge management, 2006, pp. 102–111.

[17] G. V. Cormack, T. R. Lynam, Statistical precision of information retrieval evaluation, in: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 533–540.