Preliminary Investigation on Causality Information Retrieval

Pankaj Dadure, Partha Pakray and Sivaji Bandyopadhyay

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

Abstract

Causality refers to the association of variables in a system. Humans can communicate causal interactions directly via natural language and this helps them to gain insight into how the system works. In any general context, causality is the study of the specific connection which allows the action of one event to impact others. There are several approaches that have been developed for this but still, the door is open for the inclusion of new technologies and concepts. The main focus of this work is the extraction of causal relations from unstructured text data. In which we have implemented a word embedding approach using a universal sentence encoder model has trained with a deep averaging network encoder. In which, the data of the articles are split into a new instance for each separate text body of news articles and create embeddings. For similarity measurement, cosine similarity is used. This is our primary investigation to developed the baseline system for the retrieval of causally related articles/documents.

Keywords

Causality extraction, Universal sentence encoder, Word embedding, Cosine similarity

1. Introduction

With the rapid growth and evaluation of web from where user's only able to access the information to user's are able to generate the information [1]. Thus, this process of information retrieval and generation provides the research direction for many applications like data representation, data management, data analysis, etc. In recent time, social media (like Facebook), blogging (such as Twitter) are undoubtedly well embraced technologies which mainly focused on the opinion sharing, chatting, data sharing (like photos and videos), comments, profile creation [2]. In addition to the social platform, the traditional information sharing platform like news channels, newspaper tends to be more active when it providing the feasibility of sharing, commenting, constructing, and linking documents together. These co-operative events provide the bridge for generating the data online in a huge amount.

In the current era, the automatic extraction of semantic relationships is becoming incredibly influential for questions answering, information retrieval, event prediction, generating future scenarios, and decision processing [3]. The relations for instance part-whole, if-then, cause-effect, etc shows how the event and entities recognized and compressed the pivotal information. Due to its capacity to influence decision-making, the relation between cause and effect is considered to play a very important role. However, the representation of cause-effect may vary

Forum for Information Retrieval Evaluation (FIRE)-2020, 16th-20th December, Hyderabad, India

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

and its hard to understand and formalize these variations using single style grammar. The existing cause-effect studies have numerous syntactical representations and each of them needs refinement to achieve the state-of-the-art results in a particular domain. These studies argue that the cause-effect extraction tool will have the capability to extract an undefined association between the causal and the effect component for the purpose to construct the more accurate and generic causal relation and also handle the disambiguate of any relation. The novel research direction in information retrieval is to retrieve an additional list of documents which causally relevant to search results [4]. In constraint to conventional information retrieval which typically retrieves the set of documents with respect to user's enter query, causal retrieval retrieves the set documents that describe the set of potential causes leading to an effect specified in the query. In causal retrieval system, the nature of relevance is differ from traditional topical relevance. In this paper, we have implemented the word embedding using Universal Sentence Encoder (USE) model has trained with a Deep Averaging Network (DAN) encoder.

The paper is structured as follows: Section 2 describes the prior works in domain of causal relation extraction. Section 3 gives a detailed account of the dataset. Section 4 provides detailed description about the Methodology. Section 5 describes the experimental results. Section 6 conclude with summary and directions of further research and developments.

2. Related Work

The causation knowledge extraction from natural text is under limelight over the period of time and become the significant contribution of many application like question answering [5], information retrieval [6], summarization [7], decision making [8]. The exploration of non-taxonomical relations described as the toughest task in working with the learning process. The research provides an improved context for the discovery and classification of ontological relations through a machine learning technique. For instance, in the classification of semantic relations, the meaning of the input text has taken into consideration [9]. In which initial semantic patterns through input data have drawn and extracted the patterns which assist in identifying the cause senses of input pairs of words and decides the presence of causal relations in the phrase. Dependency trees allow capturing long range relationships of words. The current approaches may ignore key details by too vigorously cutting the dependency or are insufficient to computerize because it's difficult to correlate the tree structures. For instance, graph convolutional networks are designed for the retrieval of relationships that efficiently combine knowledge on subjective dependence structures [10]. In addition, a new pruning technique to input trees has applied to add relevant information while maximally eliminating irrelevant material by holding words right around the smallest distance between two individuals in which a relationship can be formed.

A standard framework for inferences using Constrained Conditional Models (CCMs) [11], primarily described the joint challenge as an integer linear programming problem, which is consistent with time and causality limits. The joint mechanisms for inference indicate that the extraction from the document for both temporal and causal relations, is statistically significantly improved. The existing causality extraction approaches used the pattern, constraints, deep learning, and machine learning techniques which require considerable domain knowledge and

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	15

huge computation power. The BiLSTM-CRF model [12] extracts the cause and relation without analyzing the candidate causal pairs and recognized the relation independently. Moreover, the contextual string embeddings have been deployed to handle data deficiency problems and also uses the multi-head self-attention mechanism which learned the dependencies between causal words [13]. To extract the causality of social science text which contains text about the experiences of the women's life. In which applied machine learning approach has been used and extracts the causality with other relevant information. To speed-up the relationship extraction, hypothesis identification, and cause-effect entities extraction from the articles of social science [14], V. Chen et. al. developed models to classify these articles into the business and management. Moreover, the categorized hypotheses are causal relationships or not and if these are categorized as casual, then extract the cause-effect entities.

The automatic extraction of cause-effect relations using a recursive neural network [15] convey relationships in arbitrarily complex ways. The technique provides embeddings and additional linguistic characteristics to detect causal events and their consequences in a phrase. After clustering and correctly generalizing the identified events and their relationships, the causal graph is then used for predictive modeling. To extract the causality from text [16], Restricted Hidden Naive Bayes architecture considered features like contextual features, syntactic features, position features, and causal connectives. These features are extracted from the tree kernel similarity of sentences and are considered in independence. In constraint to this, restricted hidden naive bayes architecture model considered a feature in interaction and this interaction among the features avoid the over-fitting problem.

3. Dataset Description

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

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

4. Methodology

4.1. Preprocessing

4.1.1. Extraction of abbreviations and full form

Abbreviations and acronyms are commonly contained in any text article [17]. It is crucial that all types of abbreviation must be recognized in order to find the significance of abbreviations that encourage the processing of the natural language and the collection of knowledge and literature. To facilitated the abbreviation to full form mapping, we have extracted all the abbreviation and their full form from the provided the news articles and created the dictionary where keys are abbreviations and values are full forms. Afterward, all abbreviations were replaced with their respective full forms.

4.1.2. Removal of hyperlinks

For more refined and structured data, we have removed all email ids, weblink, paper references.

4.1.3. Removal Stopwords

Stopwords (I, an, is, the, etc in English) in any natural language are the most common terms [18]. These stopwords may not add enough meaning to the importance of the document when analyzing text data and constructing NLP models. Thus to save execution time and effort for processing hug amounts of text, we have removed stopwords. For this task, we have used the spaCy's inbuilt stopwords removal function.

4.1.4. Lower casing

Words like Book and book have the same syntactic and semantic meaning but when not converted to the lower case those two are represented as two different words in the vector space model[2]. To handle this, we have converted all information contains in articles into lower case.

4.1.5. Stemming

Stemming is used to make down the word (e.g. flying) to their root form (e.g. fly). In stemming, the root of the word is not the actual root word, it just a canonical form of the actual word. Stemming uses a simple heuristic method to trim the ends of terms/words in the expectation words that are correctly translated into their source [19]. So the words "benefactor", "benevolent", "beneficial" might actually be converted to "bene". There are numerous stemming algorithms. The porter Algorithm seems to be the most popular algorithm considered to be empirically useful for English.

[2] https://thehelloworldprogram.com/python/python-string-methods/

4.1.6. Converting Numbers

When the user makes a question like $10 or ten dollars. Both searched words are similar to the user's entered question. However, some IR models handle them individually and store $10 and ten dollars as different tokens. So to boost our model, we have converts 10 to ten. To achieve this we have used a num2word library[3].

4.2. Word embeddings using USE

For cause-effect extraction, the model is trained and optimized for larger length text such as text consist in provided news articles. It is trained on a variety of data like sport, business, crime, national and international issues, and attain the active responding on a wide spectrum of natural language understanding. In this trained model input is a variable length lower-cased tokenized string English text and the output is a 128-dimensional vector. The USE model [20] has trained with a DAN encoder which fragments the text body of news articles by new line into a list of the individual instance, so that embeddings is created for each text body of news articles. In DAN encoder, the embeddings of words of text body and bi-gram is averaged, then transmitted via a feed forwards deep-neural network to produce embedding for the text body. The created embeddings, generates the tensor object of shape nx128, where n is the number of text body of news articles is 303291, therefore the shape of the tensor is 303291x128. The key benefit of this model is that the computation time of the textual input is linear.

4.3. Similarity

To estimate the similarity between the text body of news articles and the user's query, the cosine similarity is taken into the consideration. The cosine similarity is used to compare the similarity between the documents and based on that it provides the ranking to the documents with respect to the user's entered query. Statistically, it calculates the cosine angle in a multidimensional space between two vectors [21].

In this work, the two vectors containing the embeddings of the text body of news articles and text based user's entered query is compared. Consider the two vector x and y. The mathematical form of cosine similarity is given as below:

$$cos(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|} = \frac{\sum_{i=0}^{n-1} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n-1} (x_i)^2} \times \sqrt{\sum_{i=1}^{n-1} (y_i)^2}} \tag{1}$$

where $\|x\|$ is the Euclidean norm of vector x=(x_1,x_2,...,x_n), defined as ${x_1}^2$+${x_2}^2$+...${x_n}^2$. Conceptually, it is the length of the vector. Similarly, $\|y\|$ is the Euclidean norm of vector y. A cosine function of 0 indicates that the two vectors have really no correlation with each other and holds the 90 degrees angle. The nearer the cosine function to 1, the lower the angle and the higher the vector match. The cosine similarity is useful, because even though the two identical texts are very distant due to their size from the Euclidean, they still have a closer angle. The angle is lower, the resemblance is higher.

[3]https://pypi.org/project/num2words/

Table 2
Obtained Results of Causality-driven Adhoc Information Retrieval Task

Run Name	MAP	P5
UCSC-Run-query_narratives	0.4553	0.7000
UCSC-Run-query_title	0.4066	0.5400
UCSC-Run-post_event_terms_expansion	0.3885	0.5000
NIT-Run-1	**0.0577**	**0.2600**

5. Experimental Results

The proficiency of the proposed approach is tested using the 15 queries which consist of a 'title' (usually a small number of keywords) and a 'narrative' (a paragraph describing the relevance criteria in detail). For each user's entered query, the proposed approach retrieves the top 40 articles based on keywords presents in the user's query. The obtained retrieved results stored in TSV file which have six fields namely query_id, iteration, document id, rank, similarity score, and run number where query_id represents the query unique identification number, iteration is fixed to 0 and it's not used by trec_tool, document id represents the unique identifier of the retrieved articles, rank attributes represents the rank of a retrieved article in a result set, score attribute represents the similarity score between the user's entered & retrieved articles and the run number represents the number of runs submitted by the participant. For evaluating the proficiency, trec_tool[4] is used, which compared the gold dataset (qrel file) with the result set obtained from the proposed system. The obtained results of article retrieval for the cause-effect relation are shown in table 2[1]. All the obtained results of the participants have been ranked based on the obtained MAP metric. For generic and more comparative analysis, the results are also estimated in terms of P@5. The obtained results for the queries "Accused Sanjay Dutt" and "Babri Masjid demolition case against Advani" are shown in table 3 and 4. As the provided query, contains the two parts i.e. title and narrative. To estimate the similarity with documents and for the retrieval, we have used the "title" part only.

6. Conclusion and Future Work

This is our primary investigation in the field of information retrieval from causally related documents. In which, we have implemented the word embedding using a universal sentence encoder model that has trained with a deep averaging network. In which, text body of the news articles are split by new instance and create embeddings for each. For similarity measurement, cosine similarity is used. This DAN based universal encoding models prepared long sentence level embeddings that shown the strong influenced in the retrieval of causally related documents. However, the proposed approach have computationally less expensive and with little lower accuracy. This depicts the room of improvement and inclusion of highly balance methodologies.

The powerful ability of feature abstraction to catch the indirect and unclear causal relations efficiently, which helps to make the majority of the current systems more accurate. So in the

[4]https://github.com/usnistgov/trec_eval

Table 3
Retrieved Results for Query "Accused Sanjay Dutt"

Sr. No.	Snippet of the retrieved articles	Articles Id
1	Sanjay Dutts producer-friend, who supplied three AK-56 rifles and a 9 mm pistol to the actor, was today sentenced to a nine-year jail term in the Bombay blasts case.	1070602_nation_story_7865037.utf8
2	The CBI believes Sanjay Dutts retraction of his confession in the Bombay blasts case may not help the actor much in court. The agencys optimism stems from the alleged similarity of Dutts version with a confession ? also retracted ? by jailed gangster Abu Salem.	1060316_nation_story_5971891.utf8
3	The fierce rivalry between underworld dons Dawood Ibrahim and Chhota Rajan tore into the high-security Arthur Road jail late on Thursday evening.	1050827_nation_story_5162970.utf8

Table 4
Retrieved Results for Query "Babri Masjid demolition case against Advani"

Sr. No.	Snippet of the retrieved articles	Article Id
1	Lal Krishna Advani and other top BJP leaders will descend on Sonia Gandhis home turf Rae Bareli on Thursday.	1050725_nation_story_5030947.utf8
2	Five of the 49 accused in the 1992 demolition of the Babri Masjid yesterday alleged that they pulled down the mosque on the instigation of Advani and other senior BJP leaders.	1030609_nation_story_2050262.utf8
3	Lalu Prasad today broke his silence on the Samajwadi Party teaming up with Kalyan Singh, a face of the Babri Masjid demolition.	1090131_nation_story_10464585.utf8

future, the recursive neural network will adopt to effectively capture implicit and ambiguous causal relations.

Acknowledgment

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

References

[1] D. Sánchez, L. Martínez-Sanahuja, M. Batet, Survey and evaluation of web search engine hit counts as research tools in computational linguistics, Information Systems, 73 (2018) pp. 50–60.

[2] S. Amer-Yahia, L. Lakshmanan, C. Yu, Socialscope: Enabling information discovery on social content sites, 4th Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, California, USA (2009) pp. 1–11.
[3] C. D. Ta, T. P. Thi, Automatic extraction of semantic relations from text documents, in: International Conference on Future Data and Security Engineering, Springer, 2016, pp. 344–351.
[4] 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.
[5] R. Girju, Automatic detection of causal relations for question answering, in: Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering, 2003, pp. 76–83.
[6] C. S. Khoo, J. Kornfilt, R. N. Oddy, S. H. Myaeng, Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing, Literary and Linguistic Computing, 13 (1998) pp. 177–186.
[7] V. Gupta, G. S. Lehal, A survey of text summarization extractive techniques, Journal of emerging technologies in web intelligence, 2 (2010) pp. 258–268.
[8] G. Loewenstein, J. S. Lerner, The role of affect in decision making, Handbook of affective science, 619 (2003) pp. 1–24.
[9] A. S. H. Al Hashimy, N. Kulathuramaiyer, Ontology enrichment with causation relations, in: IEEE Conference on Systems, Process & Control (ICSPC), IEEE, 2013, pp. 186–192.
[10] Y. Zhang, P. Qi, C. D. Manning, Graph convolution over pruned dependency trees improves relation extraction, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, Association for Computational Linguistics, (2018) pp. 2205–2215.
[11] Q. Ning, Z. Feng, H. Wu, D. Roth, Joint reasoning for temporal and causal relations, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, (2018) pp. 2278–2288.
[12] Z. Li, Q. Li, X. Zou, J. Ren, Causality extraction based on self-attentive bilstm-crf with transferred embeddings, arXiv preprint arXiv:1904.07629 (2019) pp. 1–39.
[13] R. P. Kumar, P. Aswathi, Extraction of causality and related events using text analysis, in: 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), volume 1, IEEE, 2019, pp. 1448–1453.
[14] V. Zitian Chen, F. Montano-Campos, W. Zadrozny, Causal knowledge extraction from scholarly papers in social sciences, arXiv e-prints (2020) pp. 1–12.
[15] T. Dasgupta, R. Saha, L. Dey, A. Naskar, Automatic extraction of causal relations from text using linguistically informed deep neural networks, in: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, 2018, pp. 306–316.
[16] S. Zhao, T. Liu, S. Zhao, Y. Chen, J.-Y. Nie, Event causality extraction based on connectives analysis, Neurocomputing, 173 (2016) pp. 1943–1950.
[17] H. Yu, G. Hripcsak, C. Friedman, Mapping abbreviations to full forms in biomedical articles, Journal of the American Medical Informatics Association, 9 (2002) pp. 262–272.
[18] D. Munková, M. Munk, M. Vozár, Influence of stop-words removal on sequence patterns identification within comparable corpora, in: International Conference on ICT Innovations,

Springer, 2013, pp. 67–76.
[19] A. G. Jivani, et al., A comparative study of stemming algorithms, International journal of computer technology and applications, 2 (2011) pp. 1930–1938.
[20] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al., Universal sentence encoder, arXiv preprint arXiv:1803.11175 (2018) pp. 1–7.
[21] F. Rahutomo, T. Kitasuka, M. Aritsugi, Semantic cosine similarity, in: The 7th International Student Conference on Advanced Science and Technology ICAST, volume 4, 2012, pp. 1–2.