様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 ２８ 日現在

機関番号：１４３０１

研究種目：挑戦的研究（萌芽）

研究期間：2018～2020

課題番号：１８Ｋ１９８４１

研究課題名（和文）Archive-based Question Answering

研究課題名（英文）Archive-based Question Answering

研究代表者

Ａｄａｍ　Ｊａｔｏｗｔ（Jatowt, Adam）

京都大学・情報学研究科・特定准教授

研究者番号：００４１５８６１

交付決定額（研究期間全体）：（直接経費）　4,800,000円

研究成果の概要（和文）：我々は、長期的なニュースアーカイブの質問に答えるためのアプローチを開発した。我々の手法は、過去に発行されたニュース記事からコンテンツを抽出することで、過去に関するユーザーの任意の質問に答えることができる。これは、多くの繰り返しや周期的なイベントがあるため、難しいタスクです。この研究のために、我々は答えを含む1kの質問の小さなデータセットを構築しました。我々のアプローチは、教師なしで、質問の時間範囲を推定し、その時間範囲に含まれる、または関連するニュースアーカイブからコンテンツを検索することに基づいています。

研究成果の学術的意義や社会的意義
Based on the proposed approaches users can send questions to the past and obtain detailed information without the need to manually search and browse large news article archives. Journalists, historians and anyone who wishes to obtain answers about the past can benefit from this research.

研究成果の概要（英文）：We have developed approaches for answering question in long-term news archives. Our method can answer arbitrary user query about the past by extracting content from news articles that were published long time ago. This is challenging task due to many repeating and periodical events. For this research we have built a small dataset of 1k questions that contain answers. Our approaches are unsupervised and are based on estimating question time scope and then on retrieving content from news archives that fall within or relates to that time scope. After search result reranking using special module answers are produced from individual pages and are aggregated. During the research progress we found out several important observations such as how to find the time scope in the best way or how to combine document relevance with temporal relevance of documents. We have published a paper in core A ranked conference and a journal paper that was invited from that conference submission.

研究分野：Natural language processing

キーワード：news archives　language change　question answering

１．研究開始当初の背景

Nowadays there is a lot of digitized old documents such as news article archives that can be used for extracting valuable knowledge on the past. Unfortunately, there is lack of tools dedicated for this kind of documents. Nowadays, also news is one of the most important channels for acquiring high-quality information regarding our society. However, with the rapid growth of Web, more and more news articles are available causing information overload. We proposed novel question answering techniques to alleviate users burdens with querying news archives and provide them ready answers for arbitrary questions. In addition due to the lack of large scale existing datasets for training systems to operate on news archives we have generated our own dataset and released it to the community.

２．研究の目的

The objective of this research was to design processing, search and analysis methods for answering questions over large collections of past documents such as news article archives that would be useful and attractive for either professional or average users. In addition since no existing datasets were available for questioning and understanding news archives we have prepared our own data for this purpose.

３．研究の方法

We have used natural language processing tools such as summarization, question answering techniques, as well as information retrieval or machine learning methods such as search results ranking and diversifying or word embeddings to realize our research objectives. In order to provide state of the art solutions we have incorporated novel QA frameworks such as Bertserini and BiDAF.

４．研究成果

In recent years, many old news articles have been digitized and made accessible to wide public. They serve valuable purpose in building our understanding of particular time periods in history and they preserve data about the past including information about key people, places, events, situations and etc. Different kinds of professionals (e.g., journalists, historians, sociologists) often need to deal with these collections for a variety of reasons and needs.

We proposed a large scale question answering system (QA system), which attempts to find out correct answers to questions posed in natural language over news archives. Questions about the past and also questions that could be issued to news archives tend to be usually related to particular events and exhibit certain temporal aspects. We categorized such questions into two crude types: (1) explicitly time-scoped questions: ones containing explicit temporal expressions (e.g.,“Which unarmed man was mistaken as a suspect and was shot by police in New York in 1999 ?”), and (2) implicitly time-scoped questions: ones without any explicit temporal expression in their content yet being implicitly related to specific time periods (e.g.,“Slovenia and Croatia became the first republics to declare independence from which country?”). Both types of questions resulted in different approaches to help finding their answers.

We call the large-scale question answering system that we proposed as QANA (Question Answering in News Archives). Its objective is answering the two above-mentioned types of event-related questions asked against nlong-term ews article archive collections. We note that existing QA models are mainly designed for answering questions over synchronic document collections (e.g., Wikipedia). As these systems lack the ability of utilizing temporal information, they process event-related questions and documents of the news archives in the same way as questions and documents in generic, synchronic document corpora. In contrast, QANA does not only utilize the temporal information associated with a question, but also exploits timestamp metadata of documents and the temporal information embedded in document content. Based on the combination of these kinds of temporal information it re-ranks candidate documents so as the probability of finding the correct answer in the top results is increased.

In the experimental evaluation, we tested our approach using the New York Times (NYT) Annotated corpus as a an underlying temporal document collection, based on carefully constructed test set of questions related to past events. These datasets are composed of two types of questions (explicitly and implicitly time-scoped) which have been selected from existing data sets and also from test sites focused on historical content, which makes them particularly difficult to answer. The experimental results showed that our proposed approach can improve retrieval effectiveness and surpasses the existing QA systems that are commonly used for large-scale automatic question answering.

Finally, we have proposed a framework for generating large-scale datasets for answering temporal questions in news archive collections. The lack of large-scale datasets for temporal news collections hinders the development of QA on news archives where Temporal IR techniques could be utilized. QA on historical document collections can be useful in many cases such as providing support for journalists who wish to relate their stories to certain past events, historians who investigate the past as well as employees of diverse professions, such as insurance or broad finance sectors, who wish to assess current risks based on historical accounts or support their decision making. Yes, without large datasets one cannot propose supervised approaches to Archival QA.

To overcome the shortcomings of existing QA datasets, we then devised a novel framework that assists in the creation of a diverse, large-scale ODQA dataset over a temporal document collection. The framework utilizes automatic question generation as well as a series of carefully-designed filtering steps to remove poor quality samples. As an underlying archival document collection, we used the New York Times Annotated Corpus (NYT corpus) which contains over 1.8 million news articles published between January 1, 1987 and June 19, 2007. The NYT corpus has been frequently used over the last years for many researches in temporal IR, temporal news content analysis, archival search, historical analysis and in other related tasks. The final dataset that we generated, ArchivalQA, consists of 1,067,056 data instances and is divided into different sub-parts based on the question difficulty and the containment of temporal expressions.

We chose a semi-automatic way to prepare our dataset for several reasons. First, manually generating questions would be too costly as it requires certain level of knowledge of history from annotators. Second, since question generation (QG) has recently attracted considerable attention, the available models already achieve quite good performance. Third, current ``data-hungry'' complex neural network models require larger and larger datasets to maintain good performance.

We then approached the dataset generation based on a cascade of aggressive filtering steps that remove low quality questions from a large initial pool of generated questions. We note that our dataset is not only spanning the longest time period compared to other QA datasets, but it also provides detailed questions on the events that occurred from 14 to 34 years ago. It is also one of the largest ODQA datasets available. The largest existing dataset using the temporal news collection, CNN/Daily Mail dataset has been created based on a straightforward cloze test and thus cannot be considered as a proper ODQA dataset.}.

Finally we want to say something about the potential uses of our generated dataset. Our dataset can be used in several ways. First and perhaps most commonly, QA models can use the questions, answers and paragraphs for training their IR and MRC modules Another way is to train without using the paragraph information. When it comes to the underlying news dataset, most systems would use our QA pairs against the NYT corpus. They might however use also other temporal news collections that temporally correspond to the NYT collection (i.e., ones that span 1987-2007), although naturally this will result in a more difficult task. It might be also possible to try to answer questions using synchronic knowledge bases such as Wikipedia, although as we have observed earlier, Wikipedia seems to lack a lot of detailed information on the past. The questions in our dataset are often detailed and minor and relate to old events, hence they may be different than questions in other popular ODQA datasets. Such questions can be particularly valuable considering that the true utility of QA systems lies in answering hard questions that humans cannot (at least easily) answer by themselves.

５．主な発表論文等

〔雑誌論文〕　計10件（うち査読付論文　10件／うち国際共著　10件／うちオープンアクセス　3件）

１．著者名 Jiexin Wang, Adam Jatowt, Michael Faerber, Masatoshi Yoshikawa	４．巻 1
２．論文標題 Answering Event-Related Questions over Long-Term News Article Archives	５．発行年 2020年
３．雑誌名 European Conference on Information Retrieval 2020	６．最初と最後の頁 774-789
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-45439-5_51	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Jiexin Wang, Adam Jatowt, Michael Faerber, Masatoshi Yoshikawa	４．巻 24
２．論文標題 Improving question answering for event-focused questions in temporal collections of news articles	５．発行年 2021年
３．雑誌名 Information retrieval journal	６．最初と最後の頁 29-54
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10791-020-09387-9	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Yijun Duan, Adam Jatowt, Masatoshi Yoshikawa	４．巻 DBSJ, 18(2)
２．論文標題 Comparative Summarization of Temporal Document Collections	５．発行年 2020年
３．雑誌名 Japan Society of Databases Letters (DBSJ Letters)	６．最初と最後の頁 1-6
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Yijun Duan, Adam Jatowt, and Katsumi Tanaka	４．巻 Springer, 4(4)
２．論文標題 Discovering Latent Threads in Entity Histories	５．発行年 2019年
３．雑誌名 Data Science and Engineering (DSE)	６．最初と最後の頁 336-351
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s41019-019-00108-x	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Yijun Duan, Adam Jatowt, Sourav S Bhowmick and Masatoshi Yoshikawa	４．巻 Springer, 4(3)
２．論文標題 Mapping Entity Sets in News Archives across Time	５．発行年 2019年
３．雑誌名 Data Science and Engineering (DSE)	６．最初と最後の頁 208-222
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s41019-019-00102-3	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, Nhu Van Nguyen and Antoine Doucet	４．巻 IEEE Press
２．論文標題 Post-OCR Error Detection by Generating Plausible Candidates	５．発行年 2019年
３．雑誌名 the 15th International Conference on Document Analysis and Recognition (ICDAR 2019)	６．最初と最後の頁 876-881
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICDAR.2019.00145	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Yating Zhang, Adam Jatowt, Sourav S Bhowmick, Yuji Matsumoto	４．巻 ACM Press
２．論文標題 ATAR: Aspect-based Temporal Analog Retrieval System for Document Archives	５．発行年 2019年
３．雑誌名 The 12th International Conference on Web Search and Data Mining (WSDM 2019)	６．最初と最後の頁 762-765
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3289600.3290613	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 I-Chen Hung, Michael Faeber, Adam Jatowt	４．巻 Springer LNCS
２．論文標題 Towards Recommending Interesting Content in News Archives	５．発行年 2018年
３．雑誌名 The 20th International Conference on Asia-pacific Digital Libraries (ICADL 2018)	６．最初と最後の頁 142-146
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-04257-8_13	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Yasunobu Sumikawa, Adam Jatowt	4．巻 ACM Press
2．論文標題 System for Category-driven Retrieval of Historical Events	5．発行年 2018年
3．雑誌名 The ACM/IEEE Joint Conference on Digital Libraries (JCDL 2018)	6．最初と最後の頁 413-414

掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3197026.3203888	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Adam Jatowt, Ricardo Campos, Sourav S. Bhowmick, Nina Tahmasebi, Antoine Doucet	4．巻 ACM Press
2．論文標題 Every Word has its History: Interactive Exploration and Visualization of Word Sense Evolution	5．発行年 2018年
3．雑誌名 The 27th International Conference on Information and Knowledge Management (CIKM 2018)	6．最初と最後の頁 1899-1902

掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3269206.3269218	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 Jiexin Wang
2．発表標題 Answering Event-Related Questions over Long-Term News Article Archives
3．学会等名 European Conference on Information Retrieval 2020
4．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関