A Study on Effective Approaches for Exploiting Temporal Information in News Archives(Abstract_要旨)

AUTHOR(S):

Wang, Jiexin

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	Jiexin Wang
論文題目	A Study on Effective Approaches for Exploiting Temporal Information in News Archives （ニュースアーカイブの時間情報活用のための有効な手法に関する研究）		

（論文内容の要旨）

With the application of digital preservation techniques, more and more past news articles are being digitized and made accessible online. This results in the availability of large news archives spanning multiple decades. They offer immense value to our society, contributing to our understanding of different time periods in the history and helping us to learn about the details of the past. However, the large sizes and complexities of news archives have gone far beyond user ability to utilize them efficiently. The need on how to quickly find the important, useful, precise or interesting information among an overwhelmingly large amount of news articles has rapidly arisen.

Additionally, in the news domain especially, time has long been an integral part of search engine ranking with most major search engine giving a ranking boost for recently published news articles. There are two distinct temporal aspects of a news article: timestamp (i.e., publication date) and content time (i.e., temporal expressions embedded in the document content). In the recent years, exploiting these two kinds of temporal information has been gaining increased importance in various tasks or applications, such as temporal web search, temporal question answering, search results diversification and so on.

In this dissertation, we address four research problems over temporal news collections. The first three topics investigate different approaches of incorporating two distinct temporal signals and demonstrate that the utilization of the temporal information can result in better performance or even new state-of-the-art results in various tasks. In the last topic we additionally construct a large dataset to promote the development of related research over news archives.

Chapter 1 outlines the thesis, including the research background of news archives, motivation of the research, and an overview including four research topics of the thesis.

Chapter 2 discusses two key temporal aspects of news archives and overviews previous studies related to the four research topics presented in this thesis.

Chapter 3 presents Question Answering in News Archives (QANA) system, which is designed specifically for answering two types of event-related questions on news archives. Compared with existing open-domain question

answering (ODQA) models, QANA consist of an additional module called Time-Aware Re-ranking Module, which increases the retrieval effectiveness by utilizing diverse temporal information.

Chapter 4 approaches the task of event occurrence time estimation that defined as follows: given a short event description and a chosen temporal granularity (e.g., day, week, month, or year), the task is to estimate event's occurrence time at the specified granularity. We propose TEP-Trans, which is a Transformer-based model that exploits both temporal and textual information from different angles, represented by multivariate time series. The experimental results show that TEP-Trans outperforms the existing methods by a large margin at all granularities.

Chapter 5 proposes a novel language model called TimeBERT, which is trained on a temporal news collection via two new pre-training tasks, harnessing the two kinds of temporal information to construct time-aware language representation. TimeBERT consistently outperforms BERT and other existing pre-trained models, with substantial gains on different time-related downstream tasks.

Chapter 6 introduces one of the largest ODQA datasets, called ArchivalQA, of news collections, with the object to foster the research in the field of ODQA on news archives. The dataset is constructed through a semi-automatic pipeline, whose resulting questions tend to be non-ambiguous and of good quality. In the experiments, we undertake comprehensive analysis of the generated dataset, which does not only show the quality and utility of the resulting data, but also proves the effectiveness of our question generation framework.

Finally, Chapter 7 summarizes the thesis and discusses several directions to be explored as future work.

（続紙 ２ ）

（論文審査の結果の要旨）

過去のニュース記事が大量にデジタル化され、大規模なニュースアーカイブが利用できるようになってきた。ニュースアーカイブは、歴史上の様々な時代の理解に貢献し、過去の詳細な情報を知る上で非常に大きな価値を持つ。膨大な量のニュース記事の中から、重要な情報や興味深い情報を迅速に探し出すニーズが急速に高まってきている。

本論文は、大量のニュースアーカイブから時間情報を活用する有効な手法に関する四つの課題に取り組んだ研究成果をまとめたものである。ニュース記事には、タイムスタンプ（発行日）とコンテンツタイム（文書内容に埋め込まれた時間表現）という2つの異なる時間的側面が存在する。最初の3つの課題は、これら2種類の異なる時間情報を取り入れる異なるアプローチを研究し、時間情報を利用することで、様々なタスクにおいてより良い性能、あるいは新しい結果をもたらすことができることを実証した。また、最後のトピックでは、ニュースアーカイブに関する関連研究の発展を促進するために、大規模なデータセットを構築した。具体的には、これら四つの課題の各課題について以下の成果を上げている。

第一に、質問と文書双方の多様な時間的特性を利用することで、ニュースアーカイブに対する質問に答えるための効果的なモデルを考案し、それに基づく質問応答システムを開発した。また、テストセットと20年分の文書コレクションを用いた広範な実験的評価によりその有効性を確認した。

第二に、イベントの発生時刻の推定タスクに取り組み、時間情報とテキスト情報の両方を異なる角度から利用するTransformerベースのモデルであるTEP-Transを提案した。TEP-Transモデルは、異なる粒度（例えば、日、週、月、年）で時刻を推定することができる。広範な実験を通し、TEP-Transモデルは全ての粒度で既存の手法よりも検索性能が大きく優れていることを示した。

第三に、様々な時間関連タスクに簡単かつ効果的に適用できるようにするため、TimeBERTと呼ばれる新しい言語モデルを提案した。このモデルは、2つの新しい事前学習タスクによって時間的ニュースコレクションを学習し、タイムスタンプとコンテンツタイムの2種類の時間情報を利用して時間を考慮した言語表現を構築する。TimeBERTは、時間に関するタスクにおいてBERTや他の既存の事前学習済みモデルを上回る性能を示し、イベント関連の質問に答える際の質問応答システムの性能も向上させることができることを示した。

第四に、ニュースアーカイブの質問応答分野の研究を促進する目的で、ニュースコレクションの最大級の質問応答データセットであるArchivalQAを作成した。提案した新しい質問応答データセット構築フレームワークは半自動パイプラインに基づいている。実験では、生成されたデータセットの包括的な分析を行い、生成されたデータの品質と有用性を示した。

以上、本研究の結果は、過去に蓄積された大量のニュース記事からの多角的な質問応答システムの構築に資するもので、学術上、および、実際上、寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和４年８月１８日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。また、本論文のインターネットでの全文公表についても支障がないことを確認した。

要旨公開可能日：　　年　　月　　日以降