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

平成　３０　年　　６　月　２２　日現在

機関番号：　１４３０１
研究種目：　挑戦的萌芽研究
研究期間：　2015～2017
課題番号：　１５Ｋ１２１５８

研究課題名（和文）Framework for Studying Language Evolution using Large Scale Data

研究課題名（英文）Framework for Studying Language Evolution using Large Scale Data

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）　　2,700,000円

研究成果の概要（和文）：このプロジェクトの成果には、研究方法とオンラインデモンストレーションシステムが含まれます。 我々は、時間の経過に伴う単語意味論の変化を推定するための方法とシステムを開発した。我々はまた、文書アーカイブにおける類似の用語を時間の経過とともに見つけるためのアプローチとシステムを提案した。結果はいくつかの国際的な論文に掲載されました。

研究成果の概要（英文）：The outcomes of the project involve both research methods and online demonstration systems. We have developed method and system for estimating the degree and character of semantic word evolution. We also proposed approaches and system for finding similar terms across time in document archives.

研究分野：　情報学

キーワード：　language evolution　semantic word change　across-time similarity

１．研究開始当初の背景

We have realized the lack of ready and effective methods for analyzing how terms changed over time even though word evolution is interesting to both linguistic researchers as well as wider public.

Language is constantly changing and words that have existed for long time are likely to have undergone several changes on a semantic level. As language is our most important communication tool, properly understanding the nature of changes in the meaning and usage of words – the basic elements of the language – is important for professionals working with historical texts, such as linguists, historians, librarians and social scientists. Many prior approaches relied on manual analysis of old texts. Laborious work was required to trace word occurrences across past texts and to compare their contexts and usage for drawing general inferences. As a result, manual approaches covered a relatively small set of words or time periods. Providing detailed overview of the evolution of any word over an entire timeline has however become currently possible. Computational approaches applied on large diachronic corpora have significant potential to advance evolutionary linguistic studies. Although researchers have already started proposing methods for studying word evolution, effective tools to interactively explore word change over time are largely missing or they just provide simple options such as a term frequency graph and keyword-in-context view. Equipped with effective tools, scientists would be able to freely investigate any desired word to see how it evolved over time. This could be of significant importance for those interesting in acquiring diverse kinds of knowledge reflecting word histories. Furthermore, such a tool could be useful for non-expert users, as etymology analysis attracts significant interest of the public.

Similarly, despite growing numbers of open archives containing historical documents, the search support in such document collections was inadequate. Users wanting to conduct search do not have any support in terms of query suggestions. In recent years, we have witnessed a rapid increase of text content stored in digital archives such as newspaper archives or web archives. Many old documents have been converted to digital form and made accessible online. Due to the passage of time, it is however difficult to effectively perform search within such collections. Users, especially younger ones, may have problems in finding appropriate keywords to perform effective search due to the terminology gap arising between their knowledge and the unfamiliar domain of archival collections.

２．研究の目的

The purpose of this project was to facilitate the investigation of word semantic change as well as the discovery of analogical terms over time. For the former, the target input words could be analyzed with their context represented as surrounding terms in documents containing the terms. By providing different views of word change such as semantic change, sentiment change, similarity to another reference word over time a user should be able to understand precisely how the word changed over time. Our goal was not only to deliver the framework for estimating word change but also to develop a real-time system for supporting analysis in an interactive way. The latter objective would help users searching in long-term document archives to find good keywords for their queries. Our goal was to propose a set of methods for measuring term similarity over time based on long-term document collections and an approach for explaining such a similarity by outputting some evidence.

３．研究の方法

In order to support analysis of word semantic change we have utilized large scale data resources and novel visualization methods. Google Book n-grams and Corpus of Historical American English have been used.

Several methods were proposed to investigate ways in which words evolve over time. In particular, we proposed a multi-perspective system designed for the analysis of semantic change of words and concepts over time. Using word representations from distributional semantics, we discover semantic vocabulary change and allow evolutionary word investigation at several levels: word

analysis, contrastive word pair analysis, multi-word analysis and temporal context analysis. Altogether these four core modes provide a user with a visual explanation of semantic change for any query word. Their synergy permits storytelling of the word evolution based on visual analytics. We use two datasets as underlying data: COHA and Google books n-gram dataset in order to compare the results across two datasets where one has large size while the other has balanced rate of different document genres used in each decade. Due to the very large data size of the Google Books ngrams, our methods must be scalable to be able to promptly respond to user queries.

For finding analogical terms over time we have used word embeddings and linear transformation methods. We have provided a general framework to bridge different domains across-time and, by this, to facilitate search and comparison as if carried in user's familiar domain (i.e., the present). In particular, we proposed to find analogical terms across temporal text collections by applying a series of transformation procedures. Our methods do not need any specially prepared training data and can be applied to diverse collections and time periods. We tested the performance of the proposed approaches on the collections separated by both short (e.g., 20 years) and long time gaps (70 years), and we report improvements in range of 18%-27% over short and 56%-92% over long periods when compared to state-of-the-art baselines.

We have later developed a cluster-biased transformation technique which makes use of hierarchical cluster structures built on the temporally distributed document collections. The two document datasets (one in the past and one in the present time) are subject to hierarchical clustering of terms such that term similarity is computed through comparison of their embeddings. Using the hierarchical grouping of terms, we then propose a series of transformation techniques, each one for a different cluster. The results from different transformations are later combined together to produce the final rankings of temporal analogs given a user-provided query term.

To explain term similarity across time we used systematicity principle introduced in the area of analogy and employed a random walk model on specially prepared graph. The graph contained nodes indicating pars of context terms for a given query input. The nodes were ranked based on the quality of pairs computed by random walk model over the graph. The final output was a set of context term pairs for a given input pair of similar terms. In addition, we have proposed the way to return explanatory sentences instead of terms which should be more understandable by non-professional end users.

4. 研究成果

The outcomes of the project involve both the research methods as well as online demonstration systems. For former, we developed effective approaches using word embeddings and transformation in vector spaces to develop techniques for computing term similarity in document archives. We have also proposed improved approach that utilizes cluster-based transformation techniques and a demo system for finding similar terms across time in document archives. The results were tested on both the New York Times Annotated dataset as well as on the Times digital Archive. For testing we have developed our own test sets which are available for other researchers. For word evolution analysis we proposed multi-view framework and particular pre-processing steps that should be used to efficiently and effectively analyze the way in which terms were used over time.

When it comes to demo systems, we have developed an online system for estimating the degree and character of semantic word evolution. It has precomputed data for the top 100,000 English words based on their popularity as measured by term frequency. The system can take any query input and delivers 5 different views to interactively determine the word semantic evolution. It works in real-time and also allows to download computed data in the form of csv files.

We have also constructed a demo system for finding temporal analogs. The novel points of such a system is its non-linear transformation and possibility to enter an aspect term besides a standard query term. An aspect term allows for determining the sense of a query and by this delivering more precise results. Non-linear

transformation is useful as it allows to better capture the way in which context shifted. Finally, the system also permits collecting feedback from users. Upon selecting good results by clicking on buttons next to returned analogs the user can send feedback to the system. The newly collected words are then used for extending the set of training word pairs for improving the inner-working of the system.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　13 件）

Adam Jatowt, Daisuke Kawai, Katsumi Tanaka: Temporal Analysis of Connectivity and Popularity of Historical Persons in Wikipedia, International Journal on Digital Libraries, Springer (2018)

Yasunobu Sumikawa, Adam Jatowt:
Classifying Short Descriptions of Past Events. Proceedings of the ECIR 2018: 729-736 (2018)

Yating Zhang, Adam Jatowt, Katsumi Tanaka: Temporal Analog Retrieval using Transformation based on Dual Hierarchical Structures, Proceedings of the International Conference on Information and Knowledge Management (CIKM 2017), pp. 717-726, ACM Press (2017)
DOI: 10.1145/3132847.3132917

Adam Jatowt and Ricardo Campos: Interactive System for Reasoning about Document Age, Proceedings of the International Conference on Information and Knowledge Management (CIKM 2017), demo paper, ACM Press, pp. 2471-2474 Singapore (2017)
DOI:10.1145/3132847.3133166

Adam Jatowt, Daisuke Kawai and Katsumi Tanaka: Timestamping Entities using Contextual Information, Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), ACM Press, pp. 1205-1208, Tokyo, Japan
DOI: 10.1145/3077136.3080762

Yijun Duan, Adam Jatowt and Katsumi Tanaka: Discovering Typical Histories of Entities, Proceedings of the 28th ACM Conference on Hypertext and Social Media (HT 2017), pp. 105-114, ACM Press, Prague, Czech Republic (2017)
DOI:10.1145/3078714.3078725

Yating Zhang, Adam Jatowt, Sourav S. Bhowmick, Katsumi Tanaka:
The Past is Not a Foreign Country: Detecting Semantically Similar Terms across Time. IEEE Trans. Knowl. Data Eng. 28(10): 2793-2807 (2016)

Yating Zhang, Adam Jatowt, Katsumi Tanaka: Causal Relationship Detection in Archival Collections of Product Reviews for Understanding Technology Evolution. ACM Trans. Inf. Syst. 35(1): 3:1-3:41 (2016)

Adam Jatowt, Marc Bron:
HistoryComparator: Interactive Across-Time Comparison in Document Archives. Proceedings of the COLING (Demos) 2016: 84-88 (2016)

Adam Jatowt, Daisuke Kawai, Katsumi Tanaka:
Predicting Importance of Historical Persons using Wikipedia. Proceedings of the CIKM 2016: 1909-1912 (2016)

Yating Zhang, Adam Jatowt, Katsumi Tanaka: Detecting Evolution of Concepts based on Cause-Effect Relationships in Online Reviews. Proceedings of the WWW 2016: 649-660 (2016)

Yating Zhang, Adam Jatowt, Katsumi Tanaka: Towards understanding word embeddings: Automatically explaining similarity of terms. Proceedings of the BigData 2016: 823-832 (2016)

Yating Zhang, Adam Jatowt, Sourav S. Bhowmick, Katsumi Tanaka:
Omnia Mutantur, Nihil Interit: Connecting Past with Present by Finding Corresponding Terms across Time. Proceedings of the ACL (1) 2015: 645-655 (2015)

〔学会発表〕（計　10 件）

Yasunobu Sumikawa, Adam Jatowt:
Classifying Short Descriptions of Past Events. Proceedings of the ECIR 2018: 729-736 (2018)

Yating Zhang, Adam Jatowt, Katsumi Tanaka: Temporal Analog Retrieval using

Transformation based on Dual Hierarchical Structures, Proceedings of the International Conference on Information and Knowledge Management (CIKM 2017), pp. 717-726, ACM Press (2017)

Adam Jatowt and Ricardo Campos: Interactive System for Reasoning about Document Age, Proceedings of the International Conference on Information and Knowledge Management (CIKM 2017), demo paper, ACM Press, pp. 2471-2474 Singapore (2017)

Adam Jatowt, Daisuke Kawai and Katsumi Tanaka: Timestamping Entities using Contextual Information, Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), ACM Press, pp. 1205-1208, Tokyo, Japan (2017)

Yijun Duan, Adam Jatowt and Katsumi Tanaka: Discovering Typical Histories of Entities, Proceedings of the 28th ACM Conference on Hypertext and Social Media (HT 2017), pp. 105-114, ACM Press, Prague, Czech Republic (2017)

Adam Jatowt, Marc Bron: HistoryComparator: Interactive Across-Time Comparison in Document Archives. Proceedings of the COLING (Demos) 2016: 84-88 (2016)

Adam Jatowt, Daisuke Kawai, Katsumi Tanaka: Predicting Importance of Historical Persons using Wikipedia. Proceedings of the CIKM 2016: 1909-1912 (2016)

Yating Zhang, Adam Jatowt, Katsumi Tanaka: Detecting Evolution of Concepts based on Cause-Effect Relationships in Online Reviews. Proceedings of the WWW 2016: 649-660 (2016)

Yating Zhang, Adam Jatowt, Katsumi Tanaka: Towards understanding word embeddings: Automatically explaining similarity of terms. Proceedings of the BigData 2016: 823-832 (2016)

Yating Zhang, Adam Jatowt, Sourav S. Bhowmick, Katsumi Tanaka: Omnia Mutantur, Nihil Interit: Connecting Past with Present by Finding Corresponding Terms across Time. Proceedings of the ACL (1) 2015: 645-655 (2015)

６．研究組織

(1)研究代表者
(ヤトフト アダム) Jatowt Adam
京都大学大学院 情報学研究科 社会情報学専攻特定准教授
研究者番号： 00415861