科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　　６　日現在

機関番号：３４３１５
研究種目：若手研究(B)
研究期間：2011～2013
課題番号：２３７００３０２

研究課題名（和文）古典史料からの情報抽出および可視化に関する研究

研究課題名（英文）Research of Information Extraction and Visualization from Ancient Documents

研究代表者
木村　文則（Kimura, Fuminori）

立命館大学・衣笠総合研究機構・研究員

研究者番号：７０５１６６９０

交付決定額（研究期間全体）：（直接経費）　　2,700,000円、（間接経費）　　810,000円

研究成果の概要（和文）：本研究の課題である「古典史料に対する情報抽出およびその可視化手法」を実現するには，1.古文の単語分割器の作成，2.古典史料からの知識獲得，3.得られた知見を可視化し，提示，の3つのプロセスを行う必要がある．
本研究ではこれらのプロセスに対する手法を提案し，一定の成果が得られた．上記プロセスをすべて自動化するところまでは至らなかったため，人名の抽出は人手で作成した情報を利用したものの，古文に対してテキストマイニングを行うための一連の手法を提案し，そのシステムの実装を行った．これにより，コンピュータを用いて大量の古典史料から知見を得るための支援システムを実現することができた．

研究成果の概要（英文）：I have to realize following three processes in order to achieve research of information extraction and visualization from ancient documents; 1) Word segmentation for archaic Japanese sentence, 2) Knowledge acquisition from ancient documents, and 3) Visualization of obtained knowledge.
In this research, I proposed methods for these three processes, and achieved certain result. This research has given an example of the way of information extraction and visualization from ancient documents, although I used the result of manually extraction for peoples in ancient documents because of insufficient precision of word segmentation. Therefore I realized the supporting system for obtaining knowledge from ancient documents using computers.

研究分野：総合領域

科研費の分科・細目：情報学・図書館情報学・人文社会情報学

キーワード：古典資料　情報抽出

１．研究開始当初の背景

人文科学の領域においては，古典史料を調査することにより知見を得ることは，非常に重要な研究課題である．これまでは，人文科学の研究者が多大な労力をかけ，人手により丹念に古典史料を調査することにより，研究成果を積み重ねてきた．しかし，人手による調査には，量的な限界が存在するため，調査が行き届いていない古典史料が数多く残っている．また，大量の古典史料を人手により網羅的に解析することも非常に困難である．

古典史料に対するテキストマイニングは，これまで電子化されていた古典史料が少なかったことや日本語の古文の文章を単語に分割することができなかったことなどの理由から，ほとんど研究が行われていない．しかし近年では，徐々にではあるが，古典史料が電子化されている．

これにより，電子テキスト化された古典史料の文章をコンピュータで処理することが可能となり，古典史料に対してテキストマイニングの技術を適用できる可能性がある．また，人文科学分野においてもコンピュータを用いて大量にデータを分析することが研究され始めている．古典史料に対するデータマイニングの手法の確立は，人類の過去の英知を有効に活用することであり，社会的意義があると考えられる．

２．研究の目的

本研究では，大量の古典史料に対する解析をコンピュータにより行うための手法の提案を行う．

コンピュータに古典史料を解析させるためには，古典史料が電子テキスト化されている必要がある．1990 年代半ば以降の Web の世界的な発展により，情報の電子テキスト化が活発に行われるようになり，電子化された古典史料も徐々に増加しつつある．このように，大量の古典史料に対する解析をコンピュータにより行うための環境が整い始めたことから，人文科学分野でもコンピュータを用いて古典史料を解析することが行われ始めている．

しかし残念ながら，古典史料に対するコンピュータによる解析手法はまだ研究されていない．大量の文書に対してコンピュータによる解析を行う技術として，「テキストマイニング」がある．これまでテキストマイニングは，現代語で記述された文書に対して適用されてきた．テキストマイニングの技術を古典史料にも適用できるようにすることにより，コンピュータを用いて大量の古典史料から知見を得るのが本研究の目的である．また，得られた知見を可視化することにより，古典史料を研究するための支援を行う．

３．研究の方法

本研究では，古典史料に対する情報抽出およびその可視化手法の提案を行う．本研究では，平安時代から鎌倉時代にかけて成立した古典史料を対象とする．日本語の古文で書かれたものだけでなく，『兵範記』，『吾妻鏡』などの漢文体で書かれたものも対象とする．

本研究を実現するために，以下の流れで研究を進める．

(1) 古文に対してテキストマイニングを行うために必要な古文の単語分割器の作成を行う．

(2) (1)で作成した言語資源を利用し，古典史料からの知識獲得を行う．

(3) (2)で得られた知見を可視化し，提示するシステムの構築を行う．

(1)古文の単語分割器

日本語は英語などのように単語の境界が明示されていないため，文を単語に分割することが必要となる．現代語では Chasen などの形態素解析器を用いることで行うことができるが，古文に対する形態素解析器は現在公開されていない．そこで，まず古文に対する単語分割を行う手法の研究を行う．

本研究では，古文の文章中において使用される文字の出現頻度から推定される文字 n グラムの出現頻度（理論値）と実際の文字 n グラムの出現頻度の比率から，その文字 n グラムの単語らしさを計測し，適切な単語の境界を決定する．理論値は，使用される文字の出現頻度を基に n 文字をランダムに抽出して得られる確率である．それに対し，単語を構成している文字 n グラム（適切な単語の境界が得られている文字 n グラム）は，特定の文字列を意図的に使用していることから，理論値よりも明らかに高い出現頻度が得られることになる．それゆえ，本手法では実際の出現頻度を理論値で割った値が高いほど，適切に単語の分割を行えていると仮定し，単語の分割を行う．

(2)古典史料からの知識獲得

古典史料に対しテキストマイニングを行い，知識の獲得を行う．本手法では，①古典史料から単語を抽出，②人物の特徴を生成，③人物の関連などの知識を獲得，という手順で行う．

①古典史料から単語を抽出

まず，古典史料に記述されている文章を，前年度作成した古文単語の分割器により単語に分割する．次に，構築した現代語古語対訳辞書を用いて，抽出した単語に対してラベル付けを行う．各単語に対して「人名」,「地名」,「事柄」などのラベルを付与することにより，単語の種別が判定できるようにする．このラベルは，次の人物の特徴生成において用いる単語の種類を選択する際に利用する．

②人物の特徴生成

古典史料から抽出される人物がどのような傾向があるかについての特徴を，関連する「地名」や「事柄」などを用いて表現する．まず，①において抽出された単語のうち，「人名」を取り出す．次に，その人名と共起する

地名や事柄を抽出する．抽出された共起頻度を用いて，その人物の特徴ベクトルを生成する．

③知識獲得

②で作成した人物の特徴ベクトルを用いて，人物間の特徴の類似度を求めることにより，人物間の関連や，行動の傾向の分析などを行う．

(3)獲得した知識の可視化

獲得した知識の可視化を行う．人物間の関連や人物の特徴を，ネットワーク図や 2 次元平面のグラフなどで表現し，可視化するシステムの構築を行う．

４．研究成果

(1) 古文の単語分割器

本手法の目的は，文書を単語に分割することである．本手法では，複数の異なる長さの文字 N グラムを扱うので，まず対象となる古典史料中の文章を各文字 N グラムに分割する．それらの単語らしさを評価し，その結果単語らしいと判断された文字 N グラムを単語として文の分割を行う．ここで評価する単語らしさを「単語尤度」 と呼ぶ．すなわち，「単語尤度」 が高い文字 N グラムを単語として判断する．本手法の処理手順の概要を図 1 に示す．ここで「学習データ」とは，単語の分割等がされていない単なるテキストデータであり，あらかじめ単語分割され正解を学習する目的で使用する教師データは，提案手法では必要としない．

ここで，単語である N グラムの出現頻度は，各文字の出現頻度から計算されるその N グラムの出現確率（以下「推定確率」と呼ぶ）よりもはるかに大きい頻度となるという仮定のもとで単語尤度 を定義する．「推定確率」は，文書中からランダムに n 文字抽出した際に，対象の N グラムとなる確率を意味する．この仮定より，単語尤度が高い N グラムを単語とみなすこととする．

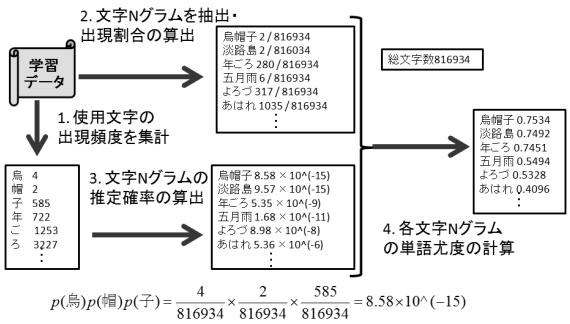

図 1．単語尤度の計算の処理の流れ

提案手法により，古文の単語分割の精度の評価を行った．「源氏物語」対象に実験を行い，「中古和文 UniDic」の分割結果を正解として評価を行った．その結果，単語分割の適合率は 0.4550，再現率は 0.5396 となった．実験結果から，実際の単語より細かく分割されてしまっていることが判明した．単語の境界について適合率は 0.6668, 再現率は 0.8199 となっており，単語境界については比較的取得できた．

	適合率	再現率	F 値
単語	0.4550	0.5396	0.4937
単語境界	0.6668	0.8199	0.7355

表 1．単語分割の精度

(2) 古典史料からの知識獲得および可視化

古典史料の本文データから，人物と地名の共起情報を用いることで人物関係の可視化を行う手法を提案した．

解析対象として古典史料のディジタルデータの一つである平安時代末期の日記である『兵範記』を用いる．人物関係の取得には，『兵範記人名索引』，地名の取得には『京都地名索引』という，人手により作成された関連史料を用いることによって，『兵範記』本文中から取得した人物と地名の共起頻度を用いる．取り出した人物と地名との共起頻度を基に，各人物の特徴をベクトルとして表現し，そのベクトルから人物間の類似度の評価を行い，人物関係図を作成する．これにより視覚的に人物関係を捉えることができる仕組みの提案を行った．また，得られた類似度を基に，K-means 法を用いて人物をクラスタリングすることで，人物のグルーピングを行った．

本手法の特徴は，人物同士の関連を直接的な関係ではなく，地名を介した間接的な関係性を基に抽出していることである．『兵範記』の書かれた平安時代では，地名と人物のつながりは現代よりも強く，地名は当時の人物の官位，家柄，役職などの特徴を表す重要な要素であるといえる．本手法は，この点を考慮した分析に適している．

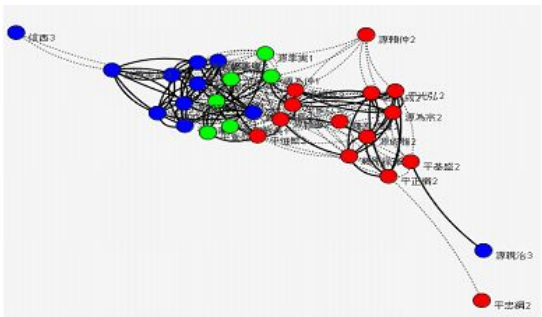

図 2．『兵範記』における保元の乱に関係する人物の関係性を可視化したグラフ

上記に示した手法で取得した人物間の関係を可視化したグラフの作成を行った．『兵範記』における保元の乱に関係する人物に対し，保元の乱がおこった 1156 年の 7 月の初めから 7 月の末までの間に絞って実験を行った（図 2）．

人物と共起した地名を，その人物特有の特徴として人物間の類似度を求めクラスタリングを行った．その結果，非常に特徴を持ったグラフの作成に成功している．クラスタ数

を 3 でクラスタリングした図 2 をみると，左右で二つに分かれ，中心の上皇派と天皇派が少し入り乱れている部分とで 3 つに分類されている．この結果は，単に上皇派と天皇派に分類するだけでなく，最前線で戦っていたと思われる第 3 のクラスタを見つけることが出来ている．この結果は，歴史的事実に沿った結果が得られると同時に，歴史学の専門家からも新たな知見が得られる可能性があると，一定の評価を受けた．

本研究の課題である「古典史料に対する情報抽出およびその可視化手法」を実現するには，以下の 3 つのプロセスを経る必要があった．

(1) 古文に対してテキストマイニングを行うために必要な古文の単語分割器の作成を行う．

(2) (1)で作成した言語資源を利用し，古典史料からの知識獲得を行う．

(3) (2)で得られた知見を可視化し，提示する．

本研究ではこれらのプロセスに対する手法を提案し，一定の成果が得られた．単語分割器の作成についてはその精度が十分とはいえないため，上記プロセスをすべて自動化するところまでは至らなかった．それゆえ，人名の抽出は人手で作成した情報を利用したものの，古文に対してテキストマイニングを行うための一連の手法を提案し，そのシステムの実装を行った．これにより，コンピュータを用いて大量の古典史料から知見を得るための支援システムを実現することができた．

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 1 件）

① 著者名：Fuminori Kimura、Takahiko Osaki、Taro Tezuka、Akira Maeda、 論文表題：Visualization of Relationships among Historical Persons from Japanese Historical Documents、雑誌名：Literary and Linguistic Computing、査読：有、Vol. 28, No. 2、発行年：2013 年、ページ：271-278、http://dx.doi.org/10.1093/llc/fqs045

〔学会発表〕（計 7 件）

① 発表者名：吉村 衛、木村 文則、前田 亮、発表表題：古文テキストからの人物表現抽出、学会名等： 人文科学とコンピュータシンポジウム、発表年月日：2013 年 12 月 13 日、発表場所：京都大学（京都府）

② 発表者名：Mamoru Yoshimura、Fuminori Kimura、Akira Maeda、発表表題：Personal Name Extraction from Ancient Japanese Texts、学会名等：The Exploration, Navigation and Retrieval of Information in Cultural Heritage ENRICH 2013 Workshop、発表年月日：2013 年 8 月 1 日、発表場所：ダブリン（アイルランド）

③ 発表者名：Mamoru Yoshimura、Fuminori Kimura、Akira Maeda、発表表題：Word Segmentation for Text in Japanese Ancient Writings Based on Probability of Character N-grams、学会名等：The 14th International Conference on Asia-Pacific Digital Libraries (ICADL2012)、発表年月日：2012 年 11 月 14 日、発表場所：台北（台湾）

④ 発表者名：吉村 衛、木村 文則、前田 亮、発表表題：古文テキスト解析のための文字 N グラムの出現確率を利用した単語分割、学会名等：人文科学とコンピュータシンポジウム、発表年月日：2011 年 12 月 11 日、発表場所：龍谷大学（京都府）

⑤ 発表者名：井坪 将、木村 文則、前田 亮、発表表題：古典史料からの相対的な人物関係の時間的変化の推定と可視化、学会名等：人文科学とコンピュータシンポジウム、発表年月日：2011 年 12 月 10 日、発表場所：龍谷大学（京都府）

⑥ 発表者名：Fuminori Kimura、Mamoru Yoshimura、Akira Maeda、発表表題：Term Extraction from Japanese Ancient Writings Using Probability of Character N-grams、学会名等：The Second International Conference on Culture and Computing (Culture and Computing 2011)、発表年月日：2011 年 10 月 22 日、発表場所：京都大学（京都府）

⑦ 発表者名：Sho Itsubo、Takahiko Osaki、Fuminori Kimura、Taro Tezuka、Akira Maeda、発表表題：Visualization of Co-occurrence Relationships Using the Historical Persons and Locational Names from Historical Documents、学会名等：Digital Humanities 2011、発表年月日：2011 年 6 月 20 日、発表場所：スタンフォード カリフォルニア（アメリカ合衆国）

６．研究組織

(1)研究代表者

木村 文則（KIMURA FUMINORI）

立命館大学・衣笠総合研究機構・研究員

研究者番号：70516690