東大寺要録からの歴史知識情報の抽出
-注釈情報の共有を目指して-

佐藤 貴文[1]　後藤 真[2]　木村 文則[3]　前田 亮[4]
[1]立命館大学 情報理工学研究科
[2]人間文化研究機構本部
[3]立命館大学 衣笠総合研究機構
[4]立命館大学 情報理工学部

人文学の分野では，ある文献に対する研究の一環として網羅的に注釈をつけることが行われている．しかし，注釈を付けるためには，様々な知識が必要となるため膨大な時間を要する．このような状況において，注釈の付与対象候補を自動的に提示することは，注釈付けの作業支援につながると考えられる．本稿では，注釈付けの作業を支援するために，日本の歴史的文書から注釈の候補になると考えられる歴史知識情報を抽出する手法を提案する．既存の注釈文字列を学習データとし，Support Vector Machine を用いて歴史知識情報とその周囲の文字とのパターンを見つけ，それを利用し，注釈の候補となると考えられる歴史知識情報を抽出する．評価実験として，k-交差検定を行い，手法の評価を行った．結果として，抽出した歴史知識情報のうち，77.93%が取り出すべき歴史知識情報と一致した．

Extracting Keyphrases for Suggesting Annotation Candidates from Japanese Historical Document “Todaiji Yoroku”
– Towards Sharing Annotation Information –

Takafumi Sato[1]　Makoto Goto[2]　Fuminori Kimura[3]　Akira Maeda[4]
[1]Graduate School of Information Science and Engineering, Ritsumeikan University
[2]The National Institutes for the Humanities (NIHU)
[3] Kinugasa Research Organization, Ritsumeikan University
[4]College of Information Science and Engineering, Ritsumeikan University

In the field of humanities, thorough annotation for a target document is carried out as research. However, the task of annotation takes an enormous amount of time because it requires a broad range of knowledge and expertise. In such a situation, automatically suggesting annotation candidates to a user will be a useful tool for supporting the task of annotation. In this paper, we propose a method for extracting keyphrases from a Japanese historical document for suggesting possible annotation candidates to annotators. We find occurrence patterns of characters surrounding existing annotations using Support Vector Machine (SVM), and extract keyphrases from non-annotated text as annotation candidates utilizing these patterns. We conducted experiments to evaluate our proposed method using k-fold cross-validation and assuming existing annotations as correct answers. As a result, we were able to extract 77.93% of correct answers as keyphrases.

１．はじめに

本論文では，『東大寺要録』という文献から歴史知識情報を抽出する手法について述べる．『東大寺要録』は，12 世紀に成立したとされる東大寺の歴史と当時の状況を記した第一級史料であり，この『東大寺要録』の情報を効果的に共有することは，日本の古代史・中世史・仏教史などの発展に重要な役割を示すものである．本論文では，漢文体で書かれた『東大寺要録』の文章を書き下した文章から歴史知識情報を抽出する手法を提案する．実験対象として，既に注釈付けと書き下し文の作成の作業が終了し，デジタル化されている『東大寺要録』の第一巻を使用する．

１．１．東大寺要録とは

『東大寺要録』は，序文によると嘉承元（1106）年に東大寺の僧が衰微した寺の再興を願って編纂したとある．しかし，実際の編者名などは知られておらず，詳細は不明である．長承 3（1134）年になって観厳が増補・再編して，現在知られる

形になっている．原本・増補本ともに散逸し，現在は，醍醐寺に第 1・第 2 巻の，東大寺に全 10 巻の写本が残されている．醍醐寺本がより古いものであるとされている．図 1 は実際の東大寺本と醍醐寺本の例である．それぞれに書き込みがしてあり，文字も異なる部分が存在する．

東大寺本は，特に東大寺における年中行事において重要な役割を果たしており，観厳の増補の後も，増補が続けられ，その宗教的・歴史的価値も非常に高いものである．現在，刊本としては，筒井英俊氏による校訂がなされた『東大寺要録』[1]，および続々群書類従に収録されているもの[2]の 2 種類がある．

現時点においては，より良好な校訂とされる筒井氏本は入手が難しくなっている．また，この両者はいずれも原漢文をそのまま文字起こししたものであり，書き下し文や詳細な注釈がついたものとはなっていない．そのため，日本史学・仏教学・建築学・美術史学などの研究者によって研究グループ（東大寺要録研究会）が組織され，注釈作業を行っている．

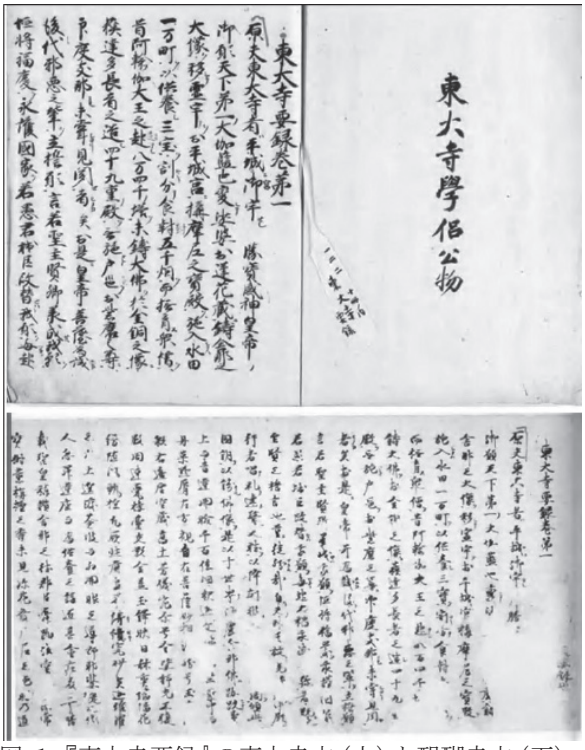

図 1:『東大寺要録』の東大寺本（上）と醍醐寺本（下）

1．2．東大寺要録のための注釈システムの現状

佐藤ら[3]は『東大寺要録』のための注釈システムを作成している．その注釈システムは，『東大寺要録』に対して注釈作業を行っている研究グループである東大寺要録研究会の研究者が使い，「複数の人物で」かつ「複雑な構造を持つ歴史的文献資料」を扱うことを想定して作成されている．そのシステムを作成するにあたって実際に使用する人文系研究者からいくつかの機能の要望を頂いている．現状のシステムでは，ユーザ登録，原文・書き下し文の閲覧，注釈の閲覧，注釈の投稿の 4 つの機能を持っている．本論文では，研究者からの要望の 1 つである，まだ注釈の作業が進んでいない本文に対して，注釈の候補となりうる文字列を提示する機能の構築について述べる．注釈の候補となりうる文字列を提示することで，注釈をつける際の一つの指標となり，注釈の作業の時間短縮に繋がると考えられる．また，このような機能を実現するために，『東大寺要録』のような日本語漢文史料から歴史知識情報を抽出することで，それらを自動で注釈の候補として提示できるようになると考えられる．本論文で対象とする歴史知識情報とは，人物表現・寺院名・仏教用語を指し，『東大寺要録』においてこれらの情報は重要であるため，注釈が付けられることが多い．また，ここでの人物表現とは，実名・別名・幼名等を含めたものである．

このような機能を実現するためには，現代日本語であれば既存の形態素解析器を適用できる．しかし，本研究で対象とする史料では，語彙や文法が異なるため，現代日本語用の形態素解析器をそのまま適用することはできない．さらに，単語への分割も困難であると考えられる．よって，歴史知識情報を抽出することは困難であると考えられる．図 2 は『東大寺要録』における歴史知識情報の例である．実際のテキストでは歴史知識情報に対するタグ付けはされていないが，ここでは説明のために赤字で表記している．

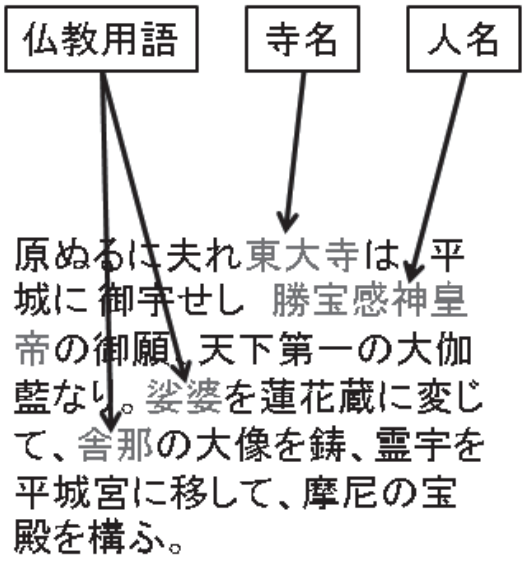

図 2『東大寺要録』における歴史知識情報の例

本論文では，教師あり機械学習手法の一つである Support Vector Machine （SVM）を用いて歴史知識情報の抽出規則を学習し，抽出を行う．SVM とは，とは教師あり学習を用いる識別手法の一つである．文章から人名や地名などの固有表現を抽出する固有表現抽出という手法において，よく使われている手法の一つに SVM がある．本手法では，SVM による固有表現抽出手法を応用して歴史知識情報の抽出を行う．

今回，漢文の原文を使わずに書き下し文を採用した理由として，実際の注釈作業では，主に書き下し文に対して注釈が付与されており，その作業に実際に役立つ機能とするためである．

本研究の最終的な目的は，抽出した歴史知識情報を用いて，古典史料における新たな注釈の候補となる箇所の提示を行う手法を確立することにある．

２．関連研究

２．１．人文学史料におけるデジタル化及び注釈付けシステムの現状

永崎ら[4][5]は『大正新脩大蔵經』という古文書のテキストデータベースの Web サービスを開発し，公開している．また，古典史料の画像及びテキストに注釈を付けるシステムとして，SMART-GS[6]がある．このように，日本の人文学史料において Web 上で古典史料の本文や注釈の閲覧や，注釈を付けるシステムは存在する．

Donato ら[7]は Pundit と呼ばれる任意の文書に対し，注釈を付けるためのシステムの構築を行っている．Pundit では，RDF トリプルを利用して文書に対して意味的な注釈を付与することができる．また，文書や既に付けられている注釈に対し，Web 上で検索したリンクを付与することも可能となっている．

しかし，注釈付けシステムにおいて歴史知識情報の注釈や，それらを注釈候補として提示するような機能に類似する機能は存在していないと考えられる．

２．２．古典テキストに対する形態素解析の試み

守岡[8][9]は電子テキスト化された古典中国語資料に対し，MeCab を用いて形態素解析を試みている．MeCab[10]とは工藤により開発された形態素解析器であり，オープンソースソフトウェアとして公開されている．MeCab は特定の言語，辞書，コーパスに依存しない汎用的な設計がされているため，辞書，コーパス，品詞体系等を用意する事により現代日本語以外の言語でも使用できる構造になっている．そこで，守岡らは古典中国語専用の辞書とコーパスを作成することにより形態素解析を行う実験を行った．

小木曽[11]らは，源氏物語や伊勢物語のような平安時代の文学作品で使われている中古和文専用の MeCab 用の辞書「中古和文 UniDic」を開発した．結果として，従来の形態素解析辞書では形態素解析することが困難だった中古和文に対し，95%を超える高い精度で解析できた．

これらの研究のように特定の時代の古文に対して適用できる形態素解析器は存在するが，古文に対して汎用的に用いることができるような形態素解析器は存在しない．同じ古文でも時代により語彙や文法が異なり，漢文体や漢字仮名交じり文といった違いもあるため，単語に分割することさえ困難な古文が存在するといった現状である．更に，今回対象とする書き下し文は，元が漢文体で書かれた文章を書き下したものであるため，どの時代の古文の文体とも合致しないといった問題が存在する．

また，形態素解析器の適用が困難な古典史料に対し，単語分割の手法を提案している研究も存在する．吉村ら[12]は文字列の出現頻度を利用した単語分割の手法を提案している．また，持橋[13]らはベイズ階層言語モデルを基とした教師なし単語分割の手法を提案している．

２．３．SVM を用いた固有表現抽出

山田ら[14]は SVM を用いた固有表現抽出の手法を提案している．単語を解析の単位とし，単語自身，品詞分類，文字種等を SVM の入力の素性として使用している．実験には CRL（郵政省通信総合研究所）固有表現データを使用している．これは，毎日新聞 95 年度版 1,174 記事，約 11,000 文に対して固有表現が付与されているものである．SVM を用いた抽出実験を行い，F 値で約 0.83 という精度を得ている．これにより SVM が固有表現抽出に有効な能力を持つとされている．この手法の問題として，形態素解析結果より短い文章の塊に対して固有表現が抽出できないといった問題が存在する．

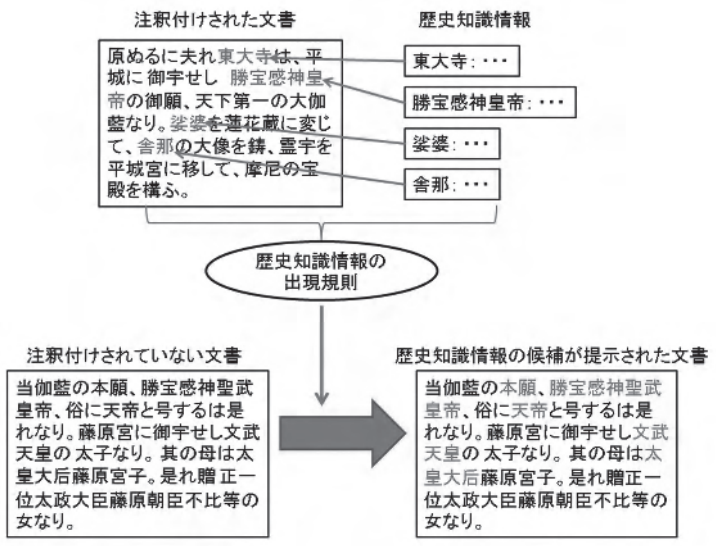

図 3：提案手法の概要図

浅原ら[15]は，この問題に対して，テキストを文字単位に分割し文字単位でまとめ上げを行う手法を提案した．文字を解析の単位とすることにより，形態素結果による分割の単語の境界と固有表現の前後の境界が一致していない場合にも対応することが可能となった．この手法では，文字単位で情報を付与するが文字には品詞情報を付与することはできない．そこで，形態素解析の品詞情報は，文字の単語内での位置情報をタグとして付与したもので代用することで表現している．

2．4．固有表現抽出の古文への適用

吉村ら[16]は，これらの SVM を用いた固有表現抽出を古文テキストに対して適用している．前述のように，特定の文体の古文テキストに対する形態素解析の研究は行われている[8][9][11]．しかし，依然として形態素解析が困難な古文テキストのほうが多い状況である．また，現代日本語の固有表現抽出では，SVM の入力として形態素解析の結果を学習・推定に利用しているため，形態素解析が困難な古文テキストにおいて，固有表現抽出ができないといった問題がある．そこで，吉村らは単語分割・素性展開・人物表現抽出の 3 つのステップからなる手法を提案し，漢文体で記述されたテキストからの固有表現抽出手法を提案した．吉村らの研究での対象のテキストは漢文体で記述された中世の日本語の史料であり，本論文で扱うような書き下し文に対しては実験を行っていない．書き下し文は，漢文体の文章とは違い，助詞や助動詞としての平仮名が含まれている．また，元の文章は漢文体なので，抽出する対象は漢字のみで構成された単語である．そのため，吉村らの手法をそのまま使用しても十分な固有表現抽出の精度が得られないと考えられる．本論文では，この精度を向上させるために，平仮名と漢字の境目を区切り情報として使用することができるのではないかと考え，この情報を取り入れた手法を提案する．

3．提案手法

本章では，SVM を用いて歴史知識情報の出現規則の学習を行い，その結果を利用し歴史知識情報の抽出を行う手法について述べる．本手法では，文章を文字単位に分割した上で機械学習を行い，機械学習の結果をまとめ上げることで歴史知識情報を抽出する．図 3 は提案手法の概要図である．注釈付けされた文書と歴史知識情報のデータから歴史知識情報の出現規則を取り出し，その出現規則を注釈付けされていない文書に適用することにより候補を抽出し，ユーザに提示する．

3．1．提案手法の構成

提案手法は学習処理と抽出処理から構成される．

3．1．1．学習処理

学習処理は以下の 3 ステップから構成される．

(1) 語の区切りの推定

入力文の語の区切りを推定する．今回対象とする書き下し文は，平仮名によって単語の区切りの推定が可能であると考えられる．また，今回は平仮名が連続した場合はそれぞれ 1 文字ずつになるように区切りの情報を挿入する．例として，『東大寺要録』の第一巻の「平城に御宇せし勝宝感神皇帝の」という箇所は，「平城/に/御宇/せ/し/勝宝感神皇帝/の/」と区切りの情報を挿入することが可能である．本論文では，歴史知識情報は人名・

寺院名・仏教用語なので全て名詞である．平仮名はほとんどが助詞・助動詞であると考えられるため，漢字で構成された部分に歴史知識情報が含まれていると考えられる．

(2) SE タグ付与

(1) の区切りの推定の結果を利用するために，Start/End（SE）というダグを使用し，前述の区切り情報に対するタグとする．これは，(1) の推定に基づき区切られた文字列が，2 文字以上の場合には先頭の文字に“B”のタグ，末尾の文字に“E”のタグ，内部の文字に“I”のタグ，1 文字になる場合には“S”のタグを付与する．表 1 は分割情報の SE タグの付与の例である．「夫れ東大寺は」を平仮名と漢字の区切りの情報を用いた分割手法により分割すると「夫/れ/東大寺/は/」となる．

(3) IOB2 タグ付与

山田ら[4]や吉村ら[7]が使用している IOB2 タグを使用し，これを歴史知識情報のタグとする．また，タグを付与する際には，文字ごとに処理する．これは固有表現の先頭のトークンに“B”，固有表現の先頭以外のトークンに“I”，固有表現以外のトークンに“O”のタグをそれぞれ付与する手法である[17]．これにより，歴史知識情報の抽出は，入力文の各トークンを IOB2 タグに分類する規則の学習であると置き換えることが可能となる．入力文の各文字に対し，IOB2 タグを歴史知識情報の出現箇所に付与する．SVM では，通常，正か負を分類する二値分類の手法であり，クラスが 3 個以上存在する場合には多値分類に拡張する必要がある．本研究では，one-versus-rest 法と呼ばれる拡張方法を用いる．これは，k 個のクラスに対し，任意のクラスかそれ以外かを分類する二値分類器を k 個構築する手法である．IOB2 タグが 3 種類存在するため，この拡張方法を用いる．表 2 は歴史知識情報の IOB2 タグ付与の例である．

(4) 歴史知識情報の学習

前の 2 ステップにより付与された IOB2 タグと SE タグを用いて SVM により出現規則のモデルデータを構築する．

表 1：区切りの推定情報の SE タグの付与の例

文字	SE タグ
夫	S-夫
れ	S-れ
東	B-東大寺
大	I-東大寺
寺	E-東大寺
は	S-は

表 2：歴史知識情報の IOB2 タグ付与の例

文字	IOB2 タグ
し	O
勝	B
感	I
神	I
皇	I
帝	I
の	O

3．1．2．抽出処理

抽出処理は以下の 3 ステップにより構成される．

(1) 語の区切りの推定

学習処理の（1）と同じ処理を行い，区切りの情報を挿入する．

(2) SE タグ付与

学習処理の（2）と同じ処理を行い，区切りの情報を用いて SE タグを付与する．

(3) 歴史知識情報の抽出

SVM により文字ごとに I，O，B のいずれかに分類し，それぞれに対応したタグを付与する．その後，付与した IOB2 タグをまとめ上げることにより，歴史知識情報の抽出を行う．SVM による分類を行う際に，学習時は IOB2 タグと SE タグの両方が既知である．その後，解析時は，SE タグの情報は既知であるが，IOB2 タグの情報は未知である．そのため，各位置で推定した IOB2 タグも次の文字の素性として利用する．推定は先頭の文字から順番に行うため，先頭から i 番目の文字に関する素性を推定する場合，i-2 番目，i-1 番目の IOB2 タグの情報も推定に用いる．「夫れ東大寺」という文の断片における「東」の推定に用いる素性を表 3 に示す．網掛けしてある項目が「東」を推定する際に導入する素性である．

図 4 は最終的なユーザへの提示の様子のイメージ図である．黄色でハイライトされている部分が抽出した歴史知識情報を表す．

表 3：使用する素性の例

位置	文字	SE タグ	IOB2 タグ
i-2	夫	S-夫	B
i-1	れ	S-れ	O
i	東	B-東大寺	
i+1	大	I-東大寺	
i+2	寺	E-東大寺	

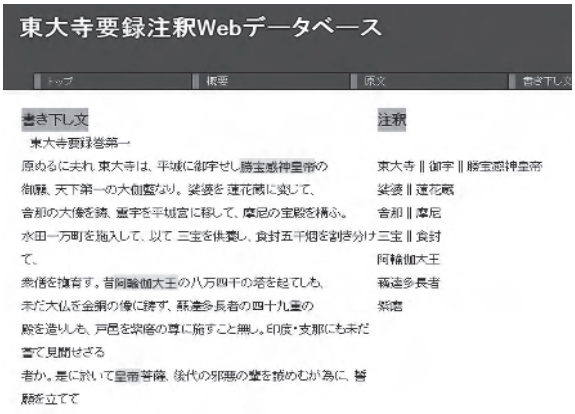

図 4：ユーザへの歴史知識情報の提示イメージ図

3．2．本手法で期待される優位性

2 章で述べた先行研究に対して，3.1 節で述べた手法を採用する理由は以下のとおりである．

a．一つの定型テキストに対して注釈を作成するための支援手法である．そのため，複数の質の異なるテキストを用いるのではなく，史料の性質がはっきりとしたものを学習して行う本手法のほうが，より精度の高い結果が期待できる．

b．本システムで対象とするテキストは原漢文とあわせて，書き下し文が作成されている．いわゆる古典テキストでもない，書き下し文をベースに注釈を作成するための手法として，本手法に優位性がある可能性がある．

c．『東大寺要録』の注釈作業が進むにつれて，再帰的に学習を行いサジェストを進めていくため，本手法の精度は注釈作業が進むほどに高まっていくと期待できる．その結果として，最終段階に近くなると注釈のつけ忘れなどを発見できるなどの，作業支援が期待できる．

4．評価実験

前章で述べた提案手法を用いて，書き下し文から歴史知識情報を抽出する実験を行った．実験データとして，『東大寺要録』の書き下し文の第一巻を使用する．正解として，既に付けられている注釈の中から人物表現・寺院名・仏教用語と考えられる単語を使用する．表 4 は使用した実験データの文字数と種類別の注釈数を示している．評価は k-交差検定と呼ばれる手法を使用する．これは，対象のデータを k 等分し，そのうちの 1 個をテストデータとし，残りの k-1 個を学習データとする手法である．学習データには既存の注釈を正解の情報として付与し学習させ，テストデータから歴史知識情報を取り出し正解と比べることで評価をする．テストデータからから抽出された歴史知識情報と正解を比べた際の適合率・再現率・F 値の平均を算出する．以下に算出式を示す．

$$\text{適合率} = (\text{正解数})/(\text{抽出した歴史知識情報数})$$

$$\text{再現率} = (\text{正解数})/(\text{抽出すべき歴史知識情報数})$$

$$\text{F 値} = (2 \times \text{適合率} \times \text{再現率})/(\text{適合率} + \text{再現率})$$

また，本論文における新たな手法として 3 章で提案した，漢字と平仮名の区切りの情報を素性として用いる手法を「提案手法」とする．また，漢字と平仮名の区切りの情報を用いない手法を「既存手法」とする．そこで，漢字と平仮名の区切りの推定情報を素性として加えない場合に対する本手法の優位性を検証する．表 5 は k=5 の場合において，提案手法と既存手法を用いた場合の実験結果である．また，表 6 は既存手法，提案手法の両方において交差検定における分割数（k）の値を変化させた実験結果である．図 5 は表 6 の適合率を，図 6 は表 6 の再現率をグラフ化した図である．それぞれ実線が提案手法による結果を表し，破線が既存手法による結果を表す．図 7，図 8 は提案手法を用いた場合の出力結果の一部である．1)は正解で，2)は提案手法により付けられた結果で，四角で囲まれた部分が歴史知識情報であると推定された箇所である．図 7 では「食封」は抽出できたが，「四天王寺」が正解だが，「四天王」までしか抽出できなかった．また，図 8 は正解ではないが，手法により抽出された歴史情報のうち，正解になる可能性のある例である．なお，抽出された「東金堂」は興福寺の中の建物の一つである．

表 4：実験データの文字数と注釈数

		『東大寺要録』第一巻書き下し文
文字数		19034
注釈数	人名	39
	寺院名	22
	仏教用語	58

表 5：提案手法と既存手法の実験結果

	適合率	再現率	F 値
既存手法	0.6647	0.6322	0.6480
提案手法	0.7462	0.5661	0.6438

表 6：分割数（k）の値を変化させた結果

		k=5	k=10	k=15	k=20
既存手法	適合率	0.6647	0.6788	0.648	0.6951
	再現率	0.6322	0.6983	0.6884	0.727
	F 値	0.648	0.6884	0.7000	0.7107
提案手法	適合率	0.7462	0.7587	0.7559	0.7793
	再現率	0.5661	0.6236	0.6494	0.6494
	F 値	0.6438	0.6845	0.6986	0.7085

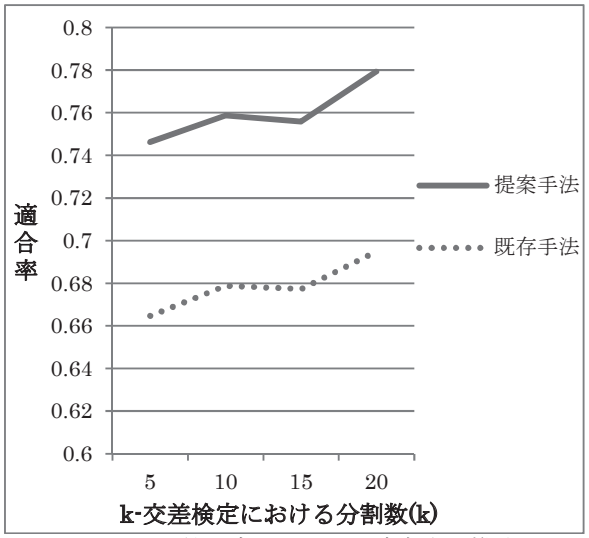

図 5：k の値の変化における適合率の推移

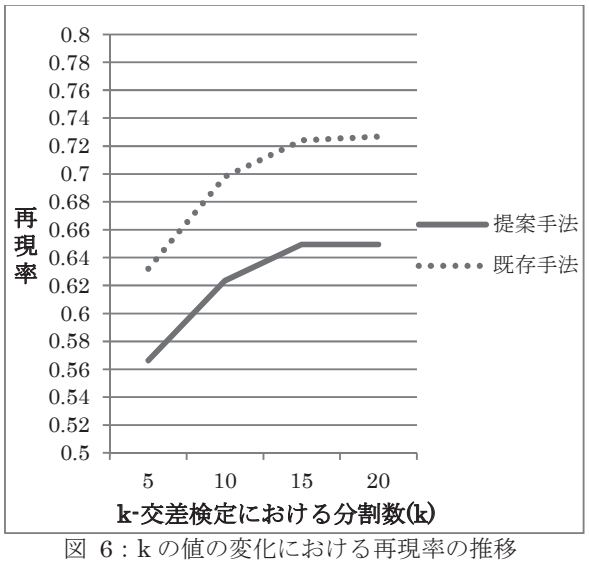

図 6：k の値の変化における再現率の推移

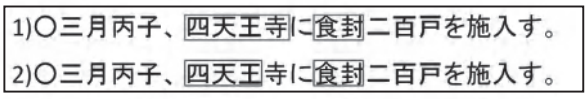

図 7：提案手法による抽出結果例

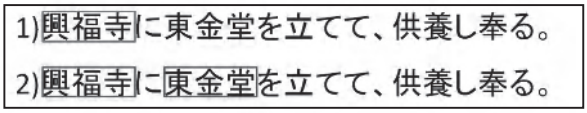

図 8：抽出された歴史知識情報が正解となる可能性のある例

5．考察

表 5 より，漢字と平仮名の区切りの情報を使用して分割した場合のほうが分割情報を使わない場合より適合率が向上した．しかし，再現率は下がっており，適合率と再現率の調和平均である F 値も下がっている．適合率は，抽出した歴史知識情報のうちどの程度正解が含まれていたかを示す．そのため，適合率が向上したことは，抽出した歴史知識情報に多く正解が含まれていたことを示す．

表 5，表 6，図 4 より，既存手法，提案手法のどちらにおいても，k-交差検定における分割数（k）の値を変化させることにより，適合率，再現率，F 値のいずれも変化した．これは，全体の分割数を多くしたことにより，学習データの量が増えたためだと考えられる．

6．おわりに

本論文では，電子テキスト化された古典史料に対して SVM を用いて歴史知識情報を抽出する手法を提案し，単語の区切り情報として漢字と平仮名の区切りの情報が素性として有効であるかどうかを検証する実験を行った．

本論文では，k-交差検定における分割数を 20 まで実験したが，今後は k をどの値にすればより良い結果が得られるかを検証する必要があると考えられる．

本手法により，新規注釈の候補を自動で推定してユーザに提示する機能を実現することができる．このような機能は既存のシステムでは実装されておらず，実際のユーザである人文系の研究者からの要望にもあることから，実現できれば有益な機能だと考えられる．

また，本論文では，既に注釈が付けられている文章に対し，交差検定を行うことで提案手法の有用性を確認したが，今後は，既に注釈が付けられている文章を訓練データとして用い，未だ注釈が付けられていない文章をテストデータとして実験することにより，注釈候補を抽出し，東大寺要録研究会のメンバーの方々に抽出した注釈が適切かどうかを判定してもらう評価実験を行うことを検討している．

謝辞

本研究に際しては，東大寺要録研究会のみなさまより忌憚なきご意見をいただくことで進められている．また，同研究会代表である大阪歴史博物館館長・東大寺史研究所所長の栄原永遠男氏のご協力をいただいている．記して感謝申し上げる．

参考文献

1) 国書刊行会:東大寺要録, 初版 1943 年, 2003 年再版.
2) 続群書類従完成会編:東大寺要録, 1969 年.
3) 佐藤貴文, 後藤真, 木村文則, 前田亮:複数の人文系研究者による史料注釈を可能とする Web システムの試作—『東大寺要録』を用いて—, 人文科学とコンピュータシンポジウム論文集, Vol.2013, No.4, pp.57-64 (2013).
4) 永崎研宣, 苫米地等流, 下田正弘:リソース連携を通じたテキスト・データベースの新たな可能性にむけて－SAT2012 を事例として－, 研究報告人文科学とコンピュータ (CH), Vol.2013-CH-97, No.1, pp.1-8 (2013).
5) 永崎研宣, 三宅真紀, 苫米地等流, A. Charles Muller, 下田正弘:人文学資料としてのテクスト構造化の意義を再考する 大正新脩大藏經における脚注の解析と Linked Data 化をめぐって, 人文科学とコンピュータシンポジウム論文集, Vol.2013, No.4, pp.239-246 (2013).
6) Yuta Hashimoto : SMART-GS Web: A HTML5-Powered, Collaborative Manuscript Transcription Platform, in Abstract of JADH 2014, pp.26-27 (2014).
7) Francesca Di Donato, Christian Morbidoni, Simone Fonda : Semantic annotation with Pundit: a case study and a practical demonstration, Proc. DH-CASE'13, ACM (2013).
8) 守岡知彦:MeCab を用いた古典中国語の形態素解析の試み, 情報処理学会研究報告. 人文科学とコンピュータ研究会報告, Vol.2008, No.73, pp.17-22 (2008)・
9) 守岡知彦:古典中国語テキストの知識処理について, Vol.85, No.1, pp.556-578 (2010)
10) 工藤拓, 山本薫, 松本裕治 : Conditional Random Fields を用いた日本語形態素解析(解析), 情報処理学会研究報告, Vol.2004, No.47, pp.89-96 (2004).
11) 小木曽智信, 小椋秀樹, 近藤明日子, 須永哲矢:中古和文を対象とした形態素解析辞書の開発, 情報処理学会研究報告, Vol.2010-CH-85, No.4, pp1-8 (2010).
12)吉村衛, 木村文則, 前田亮:古文テキストのための文字 N グラムの出現確率を利用した単語分割, Vol.2011, No.8, pp.261-268 (2011).
13) 持橋大地, 山田武士, 上田修功:ベイズ階層言語モデルによる教師なし形態素解析, 情報処理学会研究報告, Vol.2009, No.36, pp.49-56 (2009).
14) 山田寛康, 工藤拓, 松本裕治 : Support Vector Machines による日本語固有表現抽出, 情報処理学会論文誌, Vol.43, No.1, pp.44-53(2002).
15) 浅原正幸, 松本裕治:日本語固有表現抽出におけるわかち書き問題の解決, 情報処理学会論文誌, Vol.45, No.5, pp.1442-1450 (2004).
16) 吉村衛, 木村文則, 前田亮:古文テキストからの人物表現抽出, 人文科学とコンピュータシンポジウム論文集, Vol.2013, No.4, pp.97-102 (2013).
17) L. A. Ramshaw, M. P. Marcus : Text chunking using transformation-bases learning, Proc. WVLC-95, pp.83-94 (1995).