歴史資料からの人文系研究者への注釈候補の提示手法の構築

佐藤 貴文（立命館大学大学院 情報理工学研究科）
後藤 真（国立歴史民俗博物館）
木村 文則（尾道市立大学 経済情報学部）
前田 亮（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，効率的に注釈を付与するためには，様々な分野の専門知識が必要となり，膨大な時間と労力を要している．このような状況において，注釈の候補をある程度の精度で提示することが可能となれば，注釈付けの作業支援につながると考えられる．そこで我々は，日本の歴史資料から注釈を付与すべき箇所の候補を提示する手法を提案し，これを利用した注釈システムの構築を行っている．本手法では, 既存の注釈を学習データとし Support Vector Machine を用いて注釈とその周囲の文字との出現パターンを学習し，それを利用して注釈が付けられていない文章に対して注釈の候補を提示する．本論文では，対象の文章の語の区切りを推定する際の手法を改良することにより注釈候補提示の精度向上を図った．語の区切りを推定する手法として，ひらがなと漢字の区切りを利用する手法，文字 n グラムの出現確率から得られる文字連続性の学習を利用した手法，そしてひらがなと漢字の区切りを利用する手法と文字連続性の学習を利用する手法を組み合わせた手法を新たに提案する．評価実験として，提案した各提案手法の有効性の比較実験を行った．結果として，ひらがなと漢字の区切りを利用する手法と文字連続性の学習を利用した手法を組み合わせた手法でもっともよい精度が得られ，提示された注釈と既存の注釈を比較した際，84%が一致した．

A Method for Suggesting Annotation Candidates to Humanities Researchers from Japanese Historical Documents

Takafumi Sato (Graduate School of Information Science and Engineering, Ritsumeikan University)
Makoto Goto (National Museum of Japanese History)
Fuminori Kimura (Faculty of Economics, Management, and Information Science, Onomichi City University)
Akira Maeda (College of Information Science and Engineering, Ritsumeikan University)

In the research field of humanities, thorough annotation for a target document is carried out as research. However, efficient annotating for historical documents is not an easy task because it demands enormous amount of time due to the need of multiple expertise. In such a situation, suggesting annotation candidates to humanities researchers will be useful to support the task of annotation. We proposed a method for suggesting annotation candidates from a Japanese historical document to annotators in our earlier articles. The proposed method finds occurrence patterns of characters surrounding existing annotations using Support Vector Machine (SVM), and extract annotation candidates from non-annotated text utilizing these patterns. In this paper, we improve the accuracy of the proposed method by modifying the method for word segmentation of the target document. We conducted experiments to evaluate our proposed method using cross-validation. As a result, we achieved 84% of match in comparing proposed annotations and existing annotations.

1．はじめに

本論文では，『東大寺要録』という歴史資料から機械学習を用いて注釈候補を提示する手法について述べる．『東大寺要録』は，序文によると嘉承元年（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 年（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．『東大寺要

ⓒ 2015 Information Processing Society of Japan

録』には，東大寺の歴史と当時の状況を記されており，宗教的・歴史的価値が非常に高く，第一級史料に指定されている．現在，東大寺に全 10 巻の写本が残されているが，漢文体からの完全な書き下し文や詳細な注釈がついたものはない．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，書き下し作業や注釈作業を行っている．

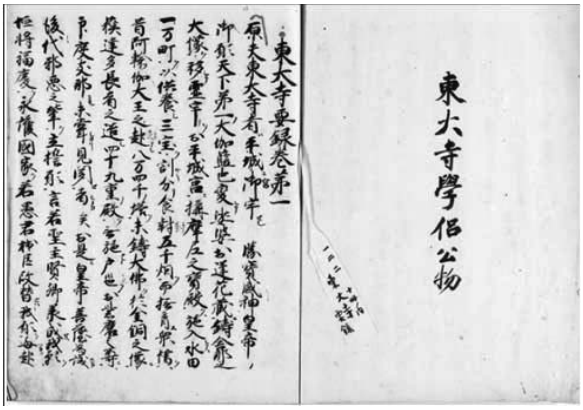

図 1：『東大寺要録』の東大寺本

現状の注釈作業の問題の一つとして，研究会のメンバーが直接顔を合わせて書き下しや注釈の検証を行っているため，膨大な時間と労力を要していることがあげられる．そこで，我々はこれらの問題を解決するために Web 上で注釈を付与することができるシステムを作成している[1]．これにより，Web 上で注釈作業を行うことができ，直接顔を合わせる前にある程度の知識の共有がされることにより時間と労力の軽減に繋がると考えている．図 2 は『東大寺要録』に付けられている注釈と，それをシステムで表示している例である．

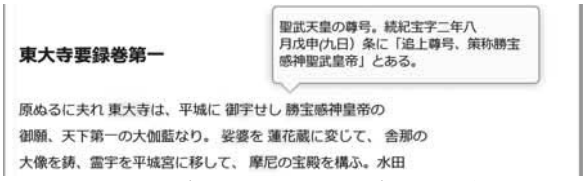

図 2：注釈システムでの注釈表示の例

また，注釈システムを作成するにあたって，実際に使用する予定の人文系研究者からいくつかの機能の要望を頂いている．その機能の一つとして，注釈の候補を自動で提示する機能があげられる．この機能を実現できれば，更に注釈作業の時間と労力を軽減することができると考えられる．そこで，我々はこの機能を実現するために，既存の注釈を学習データとし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を試みた[2]．この試みでは，『東大寺要録』の第一巻の書き下し文を対象としている．また，ここで述べている歴史知識情報とは人物表現・寺院名・仏教用語のことを指す．

本論文では，[2]での実験結果より，第一巻のうち作業が完了している書き下し文に対し，語の区切りを推定する手法を変更することにより精度の向上を試みた．[2]では語の区切りを推定する手法にひらがなと漢字の区切りを利用し，ひらがなが連続した場合，1 文字ずつ分割していた．本論文では，対象の文章の語の区切りを推定する際の手法を改良することにより注釈候補提示の精度向上を図った．語の区切りを推定する手法として，ひらがなと漢字の区切りを利用する手法，文字 n グラムの出現確率から得られる文字連続性の学習を利用した手法，そしてひらがなと漢字の区切りを利用する手法と文字連続性の学習を利用する手法を組み合わせた手法を新たに提案する．

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

Donato ら[3]は Pundit と呼ばれる任意の文書に対し注釈をつけるシステムを開発し，公開している．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，Pundit には本研究が提案するような，既存の注釈を基に注釈候補を提示する機能に類似する機能は存在していない．

２．２．SVM を用いた現代日本語に対する固有表現抽出

山田ら[4]は SVM を用いた現代日本語に対する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．

浅原ら[5]は，この問題に対してテキストを文字単位に分割し，文字ごとにそれぞれ解析した後でその結果をまとめ上げることにより固有表現を抽出している．この手法により単語分割の結果と固有表現とが一致していない場合でも SVM を用いた固有表現抽出を適用することが可能となった．

ⓒ 2015 Information Processing Society of Japan

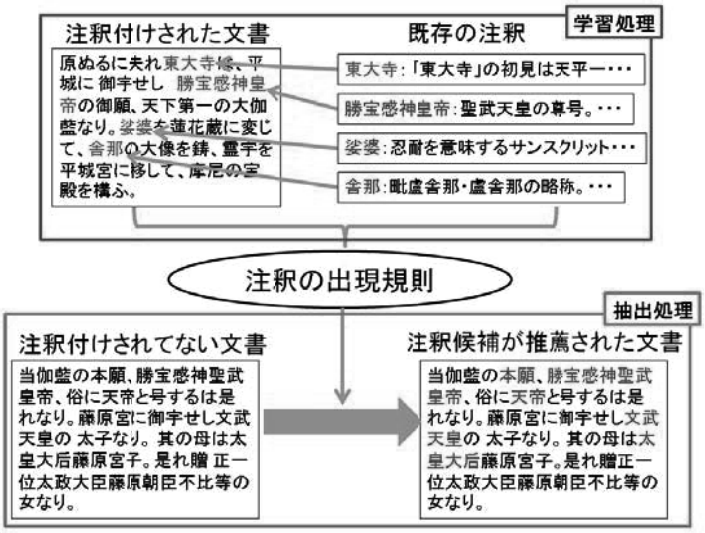

図 3: 提案手法の全体のイメージ図

２．３．SVM を用いた古文テキストに対する固有表現抽出の試み

吉村ら[6]は，SVM を用いた現代日本語に対する固有表現抽出手法[4][5]を形態素解析が困難な漢文体で書かれた古文テキストに対して適用している．吉村らの研究での対象テキストは漢文体であるがこの手法は言語や文体に依存しないため，我々は『東大寺要録』の書き下し文に吉村らの研究を応用する．また，語の区切りを推定する箇所において，書き下し文の特徴であるひらがなを利用することにより，より高い精度を得ることを目指す．

３．提案手法

本章では，人文系研究者へ提示するための注釈候補の抽出を行う手法について述べる．図 3 は提案手法の全体のイメージ図である．既に注釈が付けられた文書と既存の注釈のデータから SVM を利用し，注釈の出現規則を見つけ出す．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注釈候補を見つけ，ユーザに提示する．我々の以前の提案手法[2]では，後述の語の区切りの推定の箇所において語の区切りの推定がうまく行えていないという問題があったが，これを解決するために既存の手法におけるひらがなが連続した場合の扱いを変更した手法，文字 n グラムの出現確率から得られる文字連続性の学習を利用した分割手法，ひらがなと漢字の区切りを利用した手法に文字 n グラムの出現確率から得られる文字連続性の学習を利用した手法を組み合わせた手法を提案する．

３．１．提案手法の構成

提案手法の処理の流れを図 4 に示す．提案手法の基本的な流れは[2]の提案手法と同じである．

まず，学習処理として，対象の入力文に対して語の区切りの推定を行う．その語の区切りの推定結果を利用し，後述の SE タグを付与する．次に，対象の入力文に既存の注釈の情報を付与する．その際に後述の IOB2 タグを利用する．それらのタグの情報を利用し，SVM に学習させ，注釈の出現規則のモデルを生成する．抽出処理として，注釈が付けられていない文章に対し，語の推定を行い SE タグを付与する．その付与した文章に対し，前述のモデルを利用して注釈候補の分類を行い，抽出する．ここでは，主に変更した箇所である，学習処理における語の区切りの推定のステップと抽出処理における頻出語尾ルールの追加および部分一致箇所の判断について述べる．変更する前は，語の区切りの推定のステップにおいて，語の区切りを利用しない手法とひらがなと漢字の区切りを語の区切りに利用してひらがなが連続した場合は 1 文字ずつ分割する手法を利用していた．また，抽出処理において，頻出語尾ルールは存在せず，抽出されたあとは完全一致の判断のみを行っていた．

本提案手法では，SVM の入力の素性として，既存の注釈の情報および語の区切りの情報を利用する．その際に[2]の論文で述べた，Start/End（SE）タグと IOB2 タグを利用して解析を行う．

語の区切りの推定結果の利用には SE タグを用いる．推定結果が 2 文字以上の文字列の場合に先頭の文字には“B”，内部の文字には“I”，末尾の文字には“E”のタグをそれぞれ素性として付与する．また，推定により区切られた結果が 1 文字の場合は“S”のタグを素性として付与する．

注釈情報の利用には IOB2 タグを用いる．注釈情報の先頭の文字には“B”，注釈情報の先頭以外の文字には“I”，注釈情報でない文字には“O”のタグを素性として付与する．

これらの付与された素性を利用し注釈候補の推定を行う．学習処理では SE タグと IOB2 タグ

ⓒ 2015 Information Processing Society of Japan

の両方を既知の素性として利用し，抽出処理では SE タグのみを既知の素性として利用する．抽出処理では各文字の素性を推定しその結果をまとめ上げ抽出することにより注釈候補を抽出する．その際の推定において，対象の文字の前後 2 文字の素性を利用する．表 1 は「夫れ東大寺は」という例文における「東」の推定に用いる素性である．網掛けの箇所が「東」を推定する際に利用されている素性である．

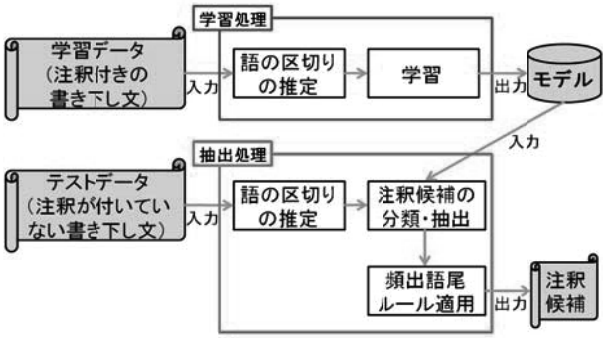

図 4：提案手法の処理の流れ図

表 1：「東」を推定する際に利用する素性

位置	文字	SE タグ	IOB2 タグ
i-2	夫	S-夫	O
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	
i+3	は	S-は	

３．２．語の区切りの推定

SVM の素性として利用するために，対象のテキストの語の区切りの推定を行う．[2]の既存手法では，ひらがなと漢字の区切りを利用して語の区切りの推定を行っていた．これは，今回対象とする書き下し文は，ひらがなと漢字が混じっており，ひらがなと漢字の境目で区切ることが可能であると考えたためである．その際に，ひらがなが連続した場合は 1 文字ずつになるように分割を行っていた．例として，「紫磨の尊に施すこと無し」という文章が入力された場合は，「紫磨/の/尊/に/施/す/こ/と/無/し/」と区切りの情報を挿入する．これは，全体の文章量が少ないため，区切りの情報を増やすことにより文字ごとの出現パターンを増加させる狙いがあった．本論文では，このステップで新たに以下に挙げる三つの手法を提案する．

３．２．１．ひらがなと漢字の区切りを利用した際にひらがなが連続した場合まとめる手法

この手法は，語の区切りの推定にひらがなと漢字の区切りを利用し，ひらがなが連続した場合にそれぞれをまとめる手法である．また，この手法を「提案手法 1」とする．既存手法では，上記で述べた理由によりひらがなをまとめずに 1 文字ずつ分割した．しかし，その分割したことにより余分な素性が付与され精度を低下させている可能性があると考えたため，ひらがなをまとめることでこれを検証する．

３．２．２．文字 n グラムの出現確率から得られる文字連続性の学習を利用した手法

この手法は，吉村ら[7]が提案した，文字 n グラムの出現確率から得られる文字連続性の学習を利用する手法である．また，この手法を「提案手法 2」とする．文字連続性を学習するため文字 n グラムの出現頻度および各文字の出現頻度から推定される出現確率を基に，その文字 n グラムの単語らしさを評価し，この単語らしさが高い文字 n グラムを単語として文章の分割を行う．ここでの単語らしさを「単語尤度」と呼ぶ．また，各文字の出現頻度から計算される n グラムの出現確率を「推定確率」と呼ぶ．ここで，単語である n グラムの出現確率は推定確率より遥かに大きくなるという仮定のもとで単語尤度を定義する．

３．２．３．提案手法 1 と提案手法 2 を組み合わせる手法

この手法は，３．２．１項の手法（提案手法 1）と３．２．２項の手法（提案手法 2）を組み合わせた手法である．この手法を「提案手法 3」とする．まず，一つ目の手法に基づき，ひらがなと漢字の境目の情報を利用した区切りを挿入する．その後，二つ目の手法に基づき，ひらがなと漢字がそれぞれ連続している箇所に対し，文字連続性の学習を利用した区切りを挿入する．これら三つの手法の分割結果の違いの例を図 5 に示す．例文は実際の『東大寺要録』の書き下し文から抜粋した．

提案手法による単語分割の違いの例
例文　　：紫磨の尊に施すこと無し
提案手法1：紫磨/の/尊/に/施/すこと/無/し/
提案手法2：紫/磨/の/尊/に/施す/こと/無し/
提案手法3：紫/磨/の/尊/に/施/す/こと/無/し/

図 5：提案手法による語の区切りの違いの例

３．３．頻出語尾ルールの追加

抽出処理において，注釈候補と判断された箇所の直後に特定の文字列が現れる場合に行うルール（頻出語尾ルール）を追加する手法である．既存の注釈を確認したところ，特定の漢字が語尾に登場することが多かったため，これらの漢字を利用したルールを適用することにより抽出精度の向上を試みた．このルールは，既存の注釈の中での登場回数が多かった，「寺」，「天皇」，「天王」，「王」のいずれかの文字列が連続した場合，

ⓒ 2015 Information Processing Society of Japan

それらの文字列まで含めて抽出というものである．例えば，「三月丙子、四天王寺に」という箇所において，SVM により「四天王」までが注釈候補であると判断された場合に，本ルールを適用することで「四天王寺」まで抽出される．

4．評価実験

前章で述べた提案手法を用いて，『東大寺要録』の書き下し文から人文系研究者へ提示するための注釈候補を抽出する実験を行った．実験データとして，『東大寺要録』の第一巻のうち作業が完了している書き下し文を使用する．本実験では，既存の注釈を，注釈が付けられるべき箇所（正解）として扱う．ただし，注釈の候補に決まった「正解」が在るわけではなく，既存の注釈が付けられていない部分でも注釈の候補として提示するのに適切なものもあると考えられるため，本実験における精度が，必ずしも求める注釈候補の精度と一致しないことに注意が必要である．本実験では，既存の注釈の中から注釈が付けられている数が多かった人物表現・寺院名・仏教用語を使用する．表 2 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．

表 2：実験データの文字数と注釈数

		『東大寺要録』書き下し文
文字数		11,249
既存の注釈数	人物表現	39
	寺院名	22
	仏教用語	58

評価は交差検定を行うことで実施する．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 個を学習データとして実験を行い，その実験をテストデータを入れ替えて k 回繰り返し平均を算出することでを評価する手法である．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，テストデータから注釈候補を取り出し「正解」と比べることで評価をする．テストデータから抽出された注釈候補と正解を比べた際の適合率・再現率・F 値を算出することにより，語の区切りを推定する手法の違いによる結果の比較を行う．以下に適合率・再現率・F 値の算出式を示す．

適合率
= 抽出した注釈候補のうち「正解」の数
/ 抽出した注釈候補の数

再現率
= 抽出した注釈候補のうち「正解」の数
/ 既存の注釈（正解）の数

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

実験には以下の手法を利用する．

1. **ベースライン**：語の区切りを用いない手法
2. **既存手法**：[2]で利用した手法であり，ひらがなと漢字の区切りを利用し，ひらがなが連続した場合 1 文字ずつ分割する手法
3. **提案手法 1**：ひらがなと漢字の区切りを利用し，ひらがなが連続した場合まとめる手法
4. **提案手法 2**：文字連続性の学習を利用した区切りを利用する手法
5. **提案手法 3**：ひらがなと漢字の区切りを使用し，ひらがなと漢字のそれぞれ連続した場合には文字連続性の学習を利用した区切りを用いた手法

また，交差検定における k の値は k=20 とする．これは，[2]の実験より k=20 の場合の実験結果が最も F 値が高かったためである．表 3 および図 6 は，それぞれの手法での抽出された注釈候補と既存の注釈が完全一致した場合の実験結果，表 4 および図 7 は，頻出語尾ルールをそれぞれの手法に適用し，注釈候補と既存の注釈が完全一致した場合の実験結果である．それぞれ，表において最大の実験結果の箇所の数値が赤字かつ太字で記されている．

実際に注釈作業をする際には，注釈候補として提示する文字列と実際に注釈を付与する箇所は必ずしも完全に一致している必要はないと考えられる．本研究における注釈候補は，注釈作業において注釈を付与する際のある程度の目安として提示するために利用されるためである．そのため，本実験により提示された注釈候補は，「正解」と一文字でも一致していれば，抽出結果として有効であると考え，抽出成功と判断した場合（部分一致）の実験を行った．この場合の部分一致とは，「正解」と抽出した注釈候補が一文字以上一致していることを指し，頻出語尾ルールは適用していない．表 5 および図 8 は，その部分一致も正解として判断した場合の実験結果である．

表 3：完全一致の場合の評価実験結果

	適合率	再現率	F 値
ベースライン	0.6933	0.6574	0.6748
既存手法	**0.8448**	0.5857	0.6918
提案手法 1	0.8382	0.5777	0.684
提案手法 2	0.7933	0.6574	0.719
提案手法 3	0.785	**0.6693**	**0.7226**

ⓒ 2015 Information Processing Society of Japan

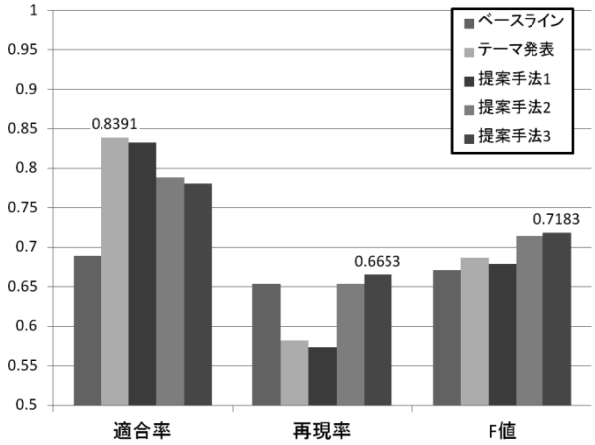

図 6：完全一致の場合の評価実験結果

表 4：頻出語尾ルールを適用し完全一致した場合の評価実験結果

	適合率	再現率	F 値
ベースライン	0.6891	0.6534	0.6708
既存手法	**0.8391**	0.5817	0.6871
提案手法 1	0.8324	0.5737	0.6792
提案手法 2	0.7885	0.6534	0.7146
提案手法 3	0.7804	**0.6653**	**0.7183**

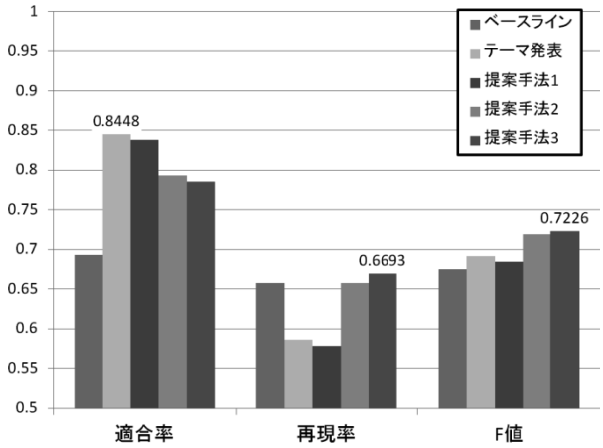

図 7：頻出語尾ルールを適用し完全一致した場合の評価実験結果

表 5：部分一致も正解に含めた場合の評価実験結果

	適合率	再現率	F 値
ベースライン	0.7395	0.7012	0.7198
既存手法	**0.8966**	0.6215	0.7341
提案手法 1	0.896	0.6175	0.7311
提案手法 2	0.8365	0.6932	0.7582
提案手法 3	0.8411	**0.7171**	**0.7742**

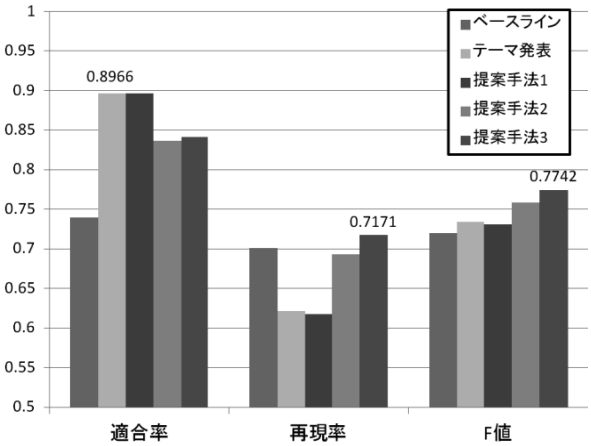

図 8：部分一致も正解と判断した場合の実験結果

5．考察

まず，ベースラインにより提示された注釈候補が含まれた文書を確認すると，語の区切りを利用していないため，助詞や助動詞といったひらがなの箇所まで含めて抽出していることが多かった．例として，「殿閣甍を連ね」と抽出されており，精度が低下する原因になっている．また，再現率が高いのは，全ての手法の中で抽出された注釈候補の数が最も多かったため再現率が最も高くなったと考えられる．しかし，抽出した候補が正解と一致している数は少なかったため適合率はすべての手法の中で最も低い結果となった．

次に，既存手法により提示された注釈候補が含まれた文書を確認すると，ベースラインに存在したような助詞や助動詞といったひらがなを含めて抽出している箇所は少なくなった．また，適合率がすべての手法の中で最も高く，抽出することができた注釈候補は「正解」と一致することが多いことがわかる．しかし，再現率が低いことが課題としてあげられる．

提案手法 1 により前述の課題を解決することが期待されたが，結果として既存手法より適合率・再現率ともに低下した結果となった．これは，ひらがなの箇所をまとめたことによりその直後に来る漢字の推定に利用する素性が減少してしまったため，精度が低下したものと考えられる．

提案手法 2 では，既存手法より適合率は 5%低下したが，再現率は約 7%向上した．本手法で提示した注釈候補は，注釈作業における目安として利用することを想定しているため，網羅的に抽出できたかを示す指標である再現率が向上することが望ましい．そのため，提案手法 2 での実験結果は望ましい結果であると考えられる．しかし，この手法は n グラムの出現確率のみを利用しており，ひらがなと漢字といった情報を利用していないため，提案手法 2 でもベースラインと同じく

ⓒ 2015 Information Processing Society of Japan

助詞や助動詞といったひらがなの箇所まで注釈候補として抽出する問題が存在する．

提案手法 3 では，提案手法 2 での問題を解決するために，提案手法 1 と同じようにひらがなと漢字の境目の情報も利用した．結果として提案手法 2 より多く注釈候補を抽出することができたため，適合率はわずかに低くなったが，再現率は全手法の中で最も高くなった．また，今回の評価実験では既存の注釈のみを「正解」として扱ったが，新たに注釈をつけるべき注釈候補がいくつか抽出されていることが提案手法 3 による実験結果において確認できた．例えば，「金鐘山寺」，「花厳経」，「甲賀寺」，「四大寺」などが注釈候補として抽出されていた．「金鐘山寺」は聖武天皇が 728 年に建立し東大寺の起源とされており，「花厳経」は華厳経という経典の別表記である．「甲賀寺」は奈良時代に存在した寺の名前であり，「四大寺」は奈良時代には大官大寺・元興寺・薬師寺・興福寺を，平安時代には，東大寺・興福寺・延暦寺・園城寺をまとめてさした呼び方である．そのため，これらは注釈候補として適切であると考えられる．

このような新たに抽出された注釈候補に関して，適切かどうか東大寺要録研究会の方に見ていただいた．結果として「正解」以外の抽出された注釈候補 22 個のうち 11 個が新たな注釈候補として適切であると判断された．このように既に注釈作業が完了している文章に対しても新たな注釈候補を提示することが可能であり，注釈作業における注釈付与の付け忘れを防ぐ効果もあると考えられる．図 9 は東大寺要録研究会の方により，適切と判断された新たな注釈候補の例である．新たな注釈候補である「金光明寺」は，金光明四天王護国之寺と呼ばれる僧寺の略称である．

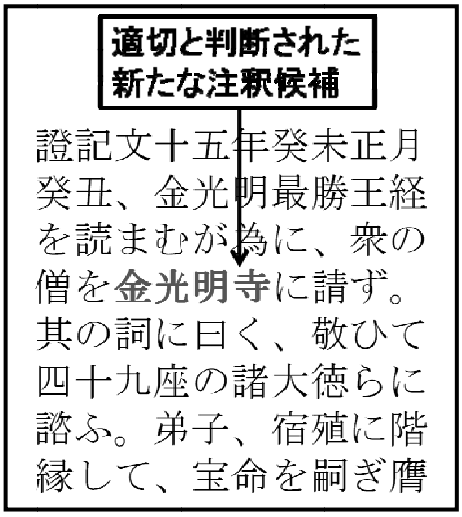

図 9：適切と判断された新たな注釈候補の例

６．おわりに

本論文では，日本語の歴史資料から人文系研究者への注釈候補を提示する手法において，既存の手法から語の区切りを推定する手法を改良することにより精度を向上させるための語の区切りを推定する手法を提案した．

今後は，より実用的な手法にするべく，注釈候補を提示する際に，その候補が注釈としてどの程度適切かの目安を確信度のような形で表示する手法を新たに検討している．その際に，既存の注釈の文字列のみを利用するのではなく，既存の注釈の説明文も利用することを検討している．

また，[1]で述べた Web 上で注釈を付与することを可能とする注釈システムの実装を完了させ，東大寺要録研究会の研究者の方々に実際に使っていただきフィードバックを得る予定である．また，現在の評価実験では，既存の注釈を，注釈が付けられるべき箇所としての「正解」と扱っているが，本論文の提案手法により提示された注釈候補が適切であるかどうかを複数の研究者に判断していただくことも検討している．

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) 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 2013，Vol. 2013，No. 4，pp.57-64（2013）．
2) 佐藤貴文，後藤真，木村文則，前田亮：『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−，人文科学とコンピュータシンポジウム論文集 2014，Vol. 2014，No. 3，pp.93-100（2014）．
3) Francesca Di Donato，Christian Morbidoni，Simone Fonda：Semantic annotation with Pundit: a case study and a practical demonstration, Proc. DH-CASE’13, ACM（2013）.
4) 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）．
5) 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）．

ⓒ 2015 Information Processing Society of Japan

6)　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 2013，Vol. 2013，No. 4，pp.97-102（2013）.
7)　吉村衛，木村文則，前田亮：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，人文科学とコンピュータシンポジウム論文集 2011，Vol. 2011，No. 8，pp. 261-268（2011）.

ⓒ 2015 Information Processing Society of Japan