古文テキストからの人物表現抽出

吉村 衛　　木村 文則　　前田 亮
立命館大学 情報理工学研究科　　立命館大学 衣笠総合研究機構　　立命館大学 情報理工学部

現在，日本語の古文テキストを解析することは，単語分割や品詞の判定が行えないため困難である．特定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．形態素解析器がない場合でも，例えば人物を表すような表現を抽出ができれば，古文テキストの解析に役立てることができる．本論文では，日本語の古文テキストから人物表現を抽出する手法を手案する．本手法では，文字の出現頻度と文字列の出現確率から単語分割を行った結果を使い，Support Vector Machine による人物表現の抽出を行った．『兵範記』，『吾妻鏡』，『玉葉』という三つの古典史料に対し実験を行った結果，単語分割情報を利用した場合，F 値で約 4%の精度の向上を確認した．

Personal Name Extraction from Ancient Japanese Texts

Mamoru Yoshimura[1]　　Fuminori Kimura[2]　　Akira Maeda[3]
[1]Graduate School of Information Science and Engineering, Ritsumeikan University
[2]Kinugasa Research Organization, Ritsumeikan University
[3]College of Information Science and Engineering, Ritsumeikan University

Text analysis of ancient Japanese language is difficult due to the lack of language tools to segment a sentence into words. There exists some morphological analysis tools for ancient Japanese in a specific period, but there are no such tools that can be used for general purpose. Even if morphological analysis tools were not available, it would be beneficial for a certain kind of text analysis to be able to extract named entities, such as personal names, from ancient Japanese texts. In this paper, we propose a method of personal name extraction from ancient Japanese texts based on Support Vector Machine (SVM) using features of character appearance and probabilistic word segmentation information. Experimental results showed that our proposed method were able to extract personal names from ancient Japanese texts with approximately 4% better F-measure when utilizing our proposed word segmentation information.

１．はじめに

近年，古文書や古記録などの古典史料が電子テキスト化されるようになってきており，その数は増加傾向にある．このことにより，現代日本語に対する自然言語処理技術を電子化された古典史料にも適用できる可能性が出てきた．現代日本語に対する自然言語処理技術では，単語の品詞特定，文章の単語への分割などを行うために形態素解析器を用いる．古典史料に対しても同様のことを行う必要があるが，現代日本語と古文としての日本語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．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も存在しない．また，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，同じ古文でも時代により語彙や文法が異なるため，それ以外の時代の日本語に対しては，単語に分割することさえ困難なのが現状である．

我々はそのような問題を解決するため，古文テキストを単語に分割する研究を行ってきた[1][2]．

形態素解析器の利用できない古文テキストから，人物を表現している部分だけでも抜き出すことができれば，古典史料に対するテキストマイニング等，古文テキストの解析に役立てることができる．たとえば，人物関係を抽出し可視化するといったことも可能となる．

そこで我々は，日本語の古文の文章から人物の表現を抽出する手法を提案している[3]．本稿では，文字の出現頻度と文字列の出現確率から単語分割を行った結果を使用し，Support Vector Machine（SVM）を用い人物表現の抽出規則を自動的に学習し，抽出を行う手法について述べる．

対象として，日本の古典史料のうち漢文体で記述されているものを使用する．なお，本論文でいう「人物表現」とは，人物の実名・別名・役職等を含めたものである．これは，今回使用するような古典史料において，人物は別名や役職で表記されていることが多いためである．漢文体の古典史料の一つである『兵範記』の一部と人物表現の例

を図 1 に示す．図では「平清盛」という人物は，名前の他に「入道」という別名，「太政大臣」という役職で表記されている．

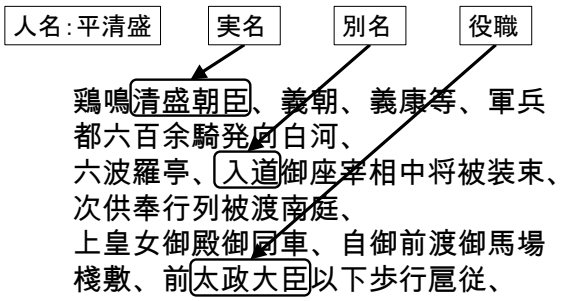

図 1　『兵範記』における人物表現の例

2．Support Vector Machine

Support Vector Machine（SVM）は教師あり学習を用いる識別手法の一つで，線形入力素子を利用して 2 クラスのパターン識別器を構成する手法である．SVM は正例・負例を分類する二値分類器であり，クラス数が 3 以上ある場合には多値分類に拡張する必要がある．代表的な拡張法として，Pairwise 法と one-versus-rest 法がある．Pairwise 法は k 個のクラスに対し任意のクラスに関する二値分類器を ${}_kC_2$ 個構築し，これらの結果の多数決により決定する手法である．one-versus-rest 法は k 個のクラスに対し，ある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．本研究では先行研究において良い結果を示している one-versus-rest 法を用いた．

今回我々は LIBLINEAR[4]という線形予測に特化した機械学習のライブラリを用いる．

3．現代日本語の固有表現抽出

機械学習を用いた固有表現抽出において，入力文を適当な解析単位（トークン）に分割し，固有表現を構成する一つもしくは複数のトークンをまとめ上げる手法が一般的である．

現代日本語の固有表現抽出において，山田ら[5]は SVM を用いた固有表現抽出を提案している．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．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．しかし，形態素解析結果の語の単位を元にしてまとめ上げると，形態素解析の単語分割と固有表現の開始・終了位置に相違があると固有表現が抽出できないという問題がある．

浅原ら[6]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って，形態素解析による分割の単語境界と，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できる．また，素性として形態素解析結果の n 次解までの品詞情報も各文字に付与している．CRL 固有表現データに対し SVM を用いた抽出実験を行い，F 値で 0.87 という結果を得ている．

上であげた研究を含め，一般的に前後 2 トークン程度の情報を素性として使用することが多い．しかし，固有表現の構成要素数が多い場合，十分な素性が与えられず，解析誤りを起こしやすい．そこで中野ら[7]は，文節区切りを行い文節内外の情報を素性として使用する手法を提案している．文節の長さに応じて使用する素性を選択し，CRL 固有表現データに対した SVM を用いた固有表現抽出の結果，F 値で 0.89 という結果を得ている．

本研究では山田ら[5]による SVM を用いた抽出手法，浅原ら[6]や中野ら[7]による文字単位での解析方法を参考に，SVM を用い文字単位で解析することで人物表現の抽出を行う．これは前述の通り，古文テキストにおいて単語に分割することが困難であり，単語を解析の単位とすることができないためである．

4．提案手法

4．1　提案手法の概要

提案手法は文字を解析の単位とし，各文字を人物表現を構成するためのラベルに分類することによって人物表現を抽出する．提案手法の処理の流れを図 2 に示す．まず，学習データ・実験データともに何らかの単語分割手法で分割を行う．その後，それぞれの文字ごとに使用する素性の付与を行い，学習データについては人物表現の正解データから正しいラベルを付与する．次に，SVM を用い学習データからモデルを構築し，実験データの各文字を分類する．最後に分類したラベルから人物表現の抽出を行う．ラベルとまとめ上げについては 4.2 節，単語分割手法については 4.3 節，利用する素性については 4.4 節で説明を行う．

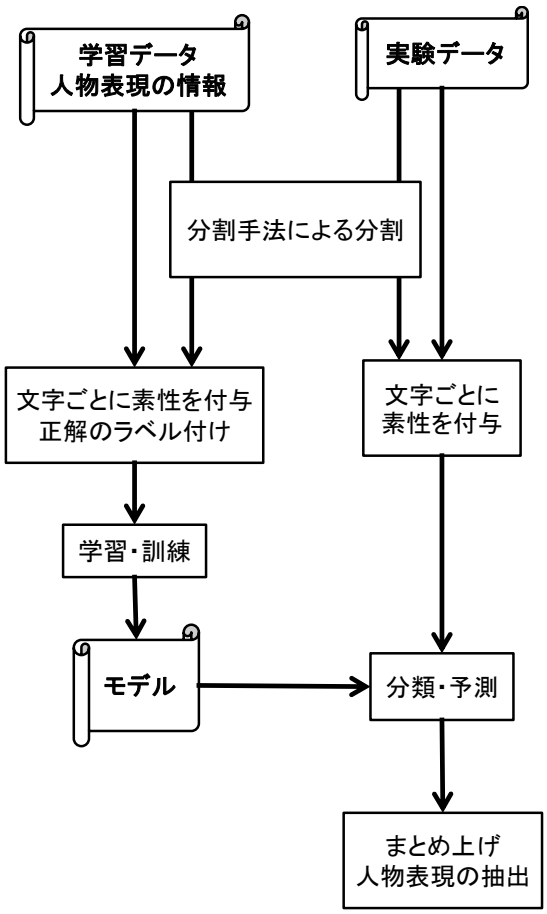

図 2　提案手法の処理の流れ

4．2 まとめ上げによる人物表現の抽出

提案手法では文字ごとに処理を行い分類していくため，最後に文字を人物表現にまとめ上げることになる．人物表現にまとめ上げる例を図 3 に示す．「上皇今朝自東山新御所」という文を文字ごとに処理していった時，最後に「上皇」という人物表現をまとめ上げることになる．

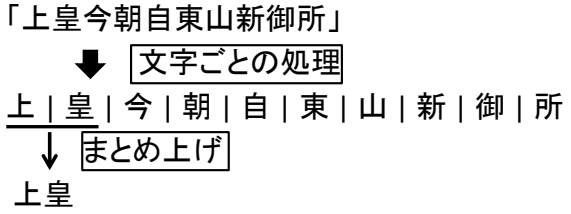

図 3　人物表現のまとめ上げ

次に，各文字を人物表現にまとめ上げる方法について説明する．分割された各トークン（文字）のチャンク（人物表現）へのまとめ上げられた状態を表現するために，タグ集合を用いる．今回は山田ら[5]の実験で最高の精度であった，IOB2 タグ集合を分類するべきラベルとして用いる．これは，チャンクの先頭のトークンに“B”のタグ，チャンクの先頭以外のトークンに“I”のタグ，チャンク以外のトークンに“O”のタグを付与する．このタグ集合を用いることにより，人物表現の抽出は，入力文の各文字を IOB2 タグのいずれかに分類する規則の学習ということができる．『兵範記』の文の一部をこの方式で IOB2 タグに分類した例を表 1 に示す．この中の「上皇」が人物表現であり，人物表現の先頭の文字「上」が B，人物表現の先頭以外の文字「皇」が I，人物表現でないそれ以外の文字が O となる．この表のように分類されたとき，B，I と並んでいる「上皇」を人物表現としてまとめ上げることができる．

表 1　IOB2 タグに分類した例

文字	IOB2 タグ
上	B
皇	I
今	O
朝	O
自	O

4．3 古文テキストへの対応

今回我々は漢文体である古文テキストを対象としている．前述の通り古文テキストに対する形態素解析が困難な為，形態素解析の結果を利用することができない．しかし，現代日本語の固有表現抽出において，形態素解析の結果の品詞・形態素の情報を SVM の入力として学習・推定に利用している．また，漢字・ひらがなといった文字種の情報も使用している．今回対象が漢文体である古文テキストであるため，これら全ての情報を利用することができない．そのため現代日本語に比べ，どうしても利用できる情報が不足してしまう．

そこで，我々がこれまでに提案した単語分割手法[1][2]で入力文を単語に分割することで，不足している情報を補う．

この手法は，文字 N グラムの単語らしさを評価し，次にこの単語らしさを用いて文の単語への分割を行う．文字 N グラムを単語の候補として扱うので，複数の異なる長さの文字 N グラムを扱う．そこで，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．

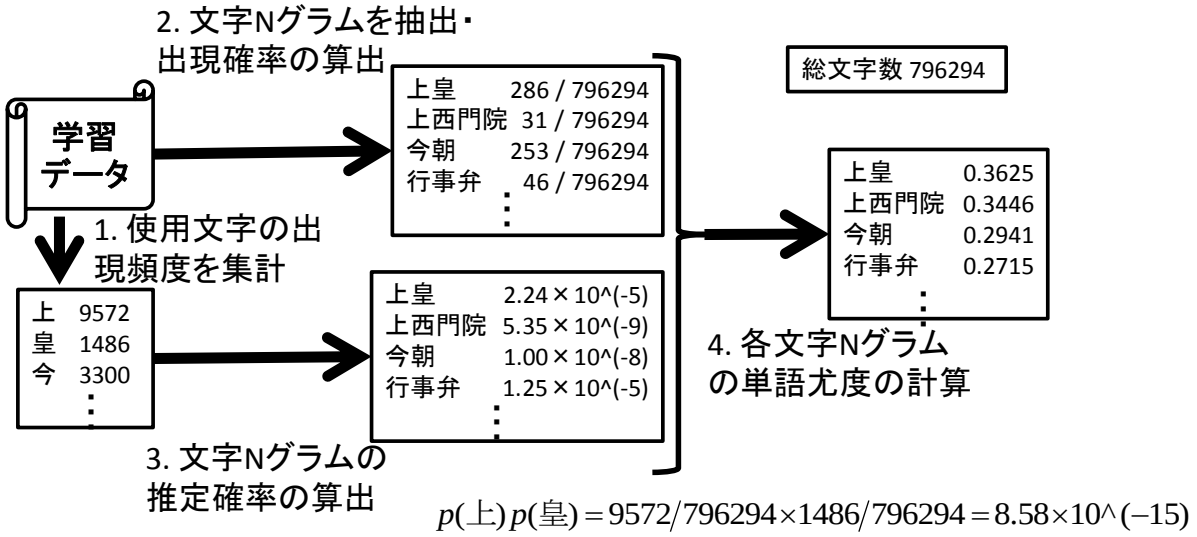

図 4　単語尤度の計算の流れ

ここで，単語である N グラムの出現確率は，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確率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．

N グラムを構成する各文字が出現する割合から求めた N グラムの推定確率と，N グラムの出現する確率との比をとることで単語尤度の計算式を定義する．

図 4 は単語尤度の計算の処理の流れを示している．まず，コーパス内の各文字の頻度を数える．次にコーパスから複数の異なる長さの文字 N グラムを抽出し，出現頻度の割合を計算する．その後，各 N グラムの推定確率を計算する．最後に単語尤度の計算を行う．

また，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，この手法では必要としない．

しかし，この手法で分割した結果は高精度というわけではないので，分割が誤りである可能性がある．つまり，分割した単語と人物表現の境界に相違がある場合がある．しかしこの手法での分割は，単語の境界での精度はある程度の値を出しており，実際の単語より細かく分割される傾向にある．そのため，我々も文字単位でまとめ上げを行うことで問題に対応する．

本手法では，文字単位で情報を利用するが，文字に直接単語情報を付与することはできない．そこで単語分割結果の情報は，分割結果に文字の単語中の位置情報を付与したものを利用する．これには，浅原ら[6]や中野ら[7]が文字単位の解析で形態素の情報を付与するために用いている Start/End（SE）タグ集合を使用する．固有表現抽出では形態素の開始・終了位置が重要な要素となるため，開始・終了位置どちらにもタグを付与する SE 法を利用する．SE 法を今回の手法に適用し，先の分割手法の分割結果に，分割結果の先頭の文字なら“B”のタグ，末尾の文字なら“E”のタグ，内部の文字なら“I”のタグ，1 文字なら“S”のタグを付与する．この SE タグを付与した分割結果を学習の素性として利用する．『兵範記』の文の一部から，この方式で SE タグを分割結果に与えた例を表 2 に示す．「上皇今朝自東山新御所」を分割手法により分割すると「上皇｜今朝｜自｜東山｜新｜御所」となる．この中の「上」は「上皇」の先頭の文字なので「B-上皇」，「皇」は末尾の文字なので「E-上皇」となる．

表 2　SE タグの付与例

文字	SE タグを付与した分割結果
上	B-上皇
皇	E-上皇
今	B-今朝
朝	E-今朝
自	S-自

4．4 使用する素性

人物表現抽出規則の学習のための素性について説明する．文頭から i 番目の文字に関する素性は，$i-2$ 番目から $i+2$ 番目までの各文字の文字自

身と単語分割結果に文字の位置を付与したもの，$i-2$ 番目と $i-1$ 番目の IOB2 タグである．「上皇今朝自・・・」という入力文における「今」に関する素性を表 3 に示す. 網掛けがしてある項目が「今」を学習する際に導入する素性である．

表 3　使用する素性の例

位置	文字	SE タグを付与した分割結果	IOB2 タグ
$i-2$	上	B-上皇	B
$i-1$	皇	E-上皇	I
I	今	B-今朝	O
$i+1$	朝	E-今朝	O
$i+2$	自	S-自	O

IOB2 タグは学習時には既知であるが，解析時は未知である．そのため各位置で推定した IOB2 タグを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，i 番目の文字を推定するときには $i-2$ 番目，$i-1$ 番目の推定された IOB2 タグがある. 例えば表 3 において i の「今」を推定する際，$i-2$ 番目の「上」と $i-1$ 番目の「皇」の推定された IOB2 タグ “B”，“I” を i 番目の「今」の推定に用いる．

５．評価実験

前節で述べた提案手法により文章から人物表現を抽出する実験を行った. 実験データとして日本語の漢文体の史料である『兵範記』,『吾妻鏡』,『玉葉』の 3 文書を使用する．これらは平安・鎌倉時代の歴史書・日記であり，電子テキスト化された本文と人手で作られた人名索引のデータがある. そのため正解データとして人名索引を利用できる. 使用する古典史料に記述されている文字数と人物表現の出現回数を表 4 に示す.

評価は『兵範記』，『吾妻鏡』，『玉葉』それぞれに対し，データを 5 等分し，訓練 4，テスト 1 の比率で交差検定を行い，それらの適合率，再現率，F 値の平均を算出する．

表 4　実験データの文字数と人物表現

	文字数	人物表現の出現回数
『兵範記』	796,294	22,488
『吾妻鏡』	787,250	39,909
『玉葉』	1,934,754	22,823

単語分割結果の情報を利用する事がどの程度精度に影響を与えているかを調べるために，単語分割結果の情報を使用した場合と使用しない場合で実験し比較を行う. 単語分割結果の情報を使用しないとは，表 3 における SE タグを付与した分割結果を素性として使わないという事である．『兵範記』，『吾妻鏡』，『玉葉』それぞれに対し行った比較実験の結果を表 5，表 6 に示す.

表 5　単語分割結果の情報を利用しない場合の結果

	適合率	再現率	F 値
『兵範記』	0.6329	0.5521	0.5897
『吾妻鏡』	0.7635	0.7220	0.7422
『玉葉』	0.6810	0.6045	0.6405

表 6　単語分割結果の情報を利用した場合の結果

	適合率	再現率	F 値
『兵範記』	0.6424	0.6162	0.6290
『吾妻鏡』	0.7971	0.7682	0.7824
『玉葉』	0.6857	0.6502	0.6786

次に，訓練・テストに用いる文書量を変更したときの抽出精度の変化を確かめるため，実験に使用する『兵範記』のテキスト量を変更し実験を行う．使用する文書量を減らし，先の実験と同じくデータを 5 等分に分割し交差検定を行った. データ量以外の条件は同じである. 結果を図 5 に示す.

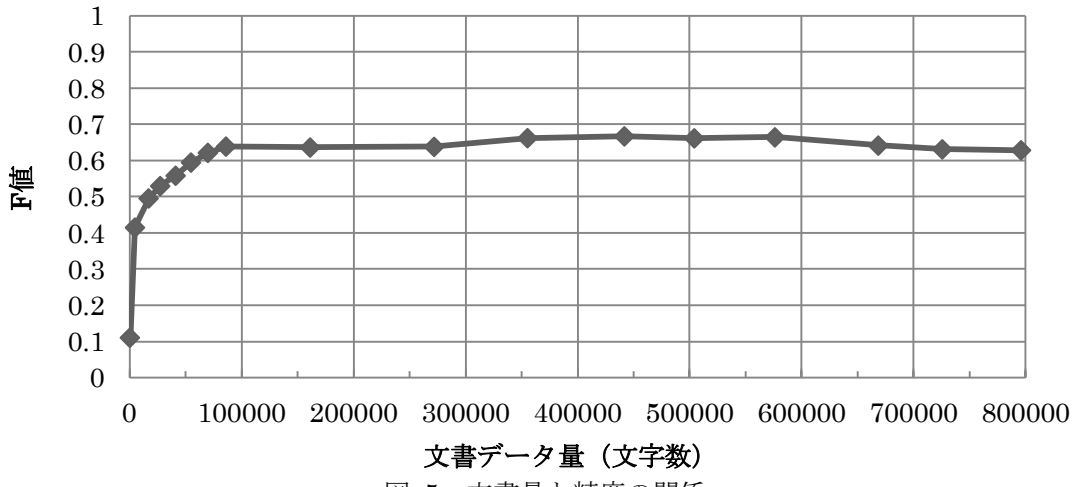

図 5　文書量と精度の関係

６．考察

表 5,6 から，3 文書とも単語分割結果の情報を素性として利用した方が精度が高くなることがわかった．

図 5 の使用する文書量と抽出精度の関係の実験結果をみると，文字数が増えるにつれ F 値が上昇していき，途中から横ばいになっていることがわかる．この横ばいになっている部分が 8 万文字あたりからである．5 等分の交差検定を行っているので，この手法では学習データとして約 6.4 万文字あれば十分な結果が得られることがわかる．この研究は古文テキストの解析を目的としているが，古典史料はデータ量に限りがあるため，大量のデータを用意することができない可能性がある．よって，どれだけ少ないデータ量で十分な結果が得られるかということも重要な指標である．

７．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，Support Vector Machine を用いて人物表現を抽出する手法を提案した．形態素解析が難しく利用できる情報が少ない古文テキストに対して，単語分割手法を使うことによって得た分割結果を素性として用い精度の向上を図った．

今回は同文書内でのみの実験を行ったが，今後は異なる文書間での実験も行う必要がある．例えば『兵範記』で学習を行い，『吾妻鏡』の人物表現を抽出するといったことである．なぜなら，人物表現のデータがない史料に対して，どれほど正確に人物表現を抽出できるかを実験する必要があるためである．

また，今後はさらなる精度の向上を目指す必要がある．現代日本語に比べ利用できる情報が少ないため，高い精度を出すことが難しくなっている．そこで情報を少しでも補強するために，新たに辞書のデータを用いる事を検討している．例えば名詞と特定できる一部文字列について限定的に情報を与えることにより，精度が向上できると考えている．表 3 の例であれば，辞書により「上皇」が名詞であると特定できれば，「今」に前の文字が名詞という素性を付与できる．

謝辞

本研究の一部は文部科学省私立大学戦略的研究基盤形成支援事業「芸術・文化分野の資料デジタル化と活用を軸とした研究資源共有化研究」，文部科学省科学研究費補助金若手研究(B)「古典史料からの情報抽出および可視化に関する研究」(研究代表者:木村文則，課題番号:23700302)，文部科学省科学研究費補助金基盤研究（C）「多言語ディジタルアーカイブの統合検索に関する研究」（研究代表者 : 前田亮，課題番号 : 24500300）の支援を受けている．

参考文献

1) Mamoru Yoshimura, Fuminori Kimura, Akira Maeda. Word Segmentation for Text in Japanese Ancient Writings Based on Probability of Character N-grams, In Proc. ICADL2012, pp. 313-316, 2012.

2) 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, pp. 261-268, 2011.

3) Mamoru Yoshimura, Fuminori Kimura, and Akira Maeda. Personal Name Extraction from Ancient Japanese Texts, In Proc. ENRICH2013 Workshop, pp. 31-34, 2013.

4) Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9, pp. 1871-1874, 2008.

5) 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.

6) 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.

7) 中野桂吾，平井有三: 日本語固有表現抽出における文節情報の利用，情報処理学会論文誌, Vol.45, No.3, pp.934-941, 2004.