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

吉村 衛　　木村 文則　　前田 亮
立命館大学 情報理工学部

現在，日本語の古文に対して汎用的に用いることができる形態素解析器は存在しない．それゆえ日本語の古文に対しては，文章を単語に分割することさえ困難である．単語分割が行えるようになると，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，日本語の古文の文章を単語に分割する手法を手案する．本手法では，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の単語への分割を行う．今回は，「源氏物語」に対し本手法の評価実験を行い，評価・考察を行う．

Term Extraction for Text Analysis of Japanese Ancient Writings Based on Probability of Character N-grams

Mamoru Yoshimura　Fuminori Kimura　Akira Maeda
College of Information Science and Engineering,
Ritsumeikan University

Currently, there are few available tools to separate ancient Japanese sentences into terms. Therefore, it is difficult to extract archaic Japanese terms from Japanese ancient writings. In this paper, we propose a method of term extraction for Japanese ancient writings. We calculate the likelihood of character n-grams to be a term, and extract character n-grams with higher likelihood as archaic Japanese terms. We conducted experiments of term separation using the term likelihood by the proposed method.

1. まえがき

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．また，特定の時代の日本語を対象とした形態素解析用の辞書は存在するが，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．古文の単語分割が行えるようになると，まず古典史料用の辞書の作成に役立てることができる．また，人名や地名が抽出できると，そこから人物関係を検出することも可能となる可能性がある．よって，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．

そこで我々は，日本語の古文の文章を単語分割する手法を提案する．本手法では，文字 N グラムの出現頻度および理論上推定されるその出現確率を基に，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の分割を行う．本手法では単語らしさを求める際に，辞書などの言語資源を必要としないことが特徴である．本稿では日本語の古文の解析を目的としているが，手法自体は特定の言語に依存しないため，様々な時代，言語に適用が可能なことも本手法の特徴と言える．

2. 関連研究

古典史料を解析できるようにした形態素解析辞書に「中古和文 UniDic」[1]がある．これは特定の時代，言語のみに対応したもので，本手法とは利用可能な範囲が大きく違う．倉田ら[2]は単語の区切りを検出するため，日本語コーパスを用いて確率的言語モデルを適用した．この手法では，単語 N グラム確率は，単語列の N グラムの頻度の割合を，(N-1)グラムの頻度の割合で割ることによって得られる．また，持橋ら[3]は教師なし単語分割のため，ベイズ階層言語モデルを提案している．倉田ら[2]の手法では，単語 N グラムの確率を推定するための学習処理，持橋ら[3]の手法ではギブスサンプリングを用いた学習処理のコストがかかる．これに対し我々の提案手法ではこの学習処理を必要とせず，比較的シンプルな手法で実用的な処理速度および分割精度を実現することを目標としている．

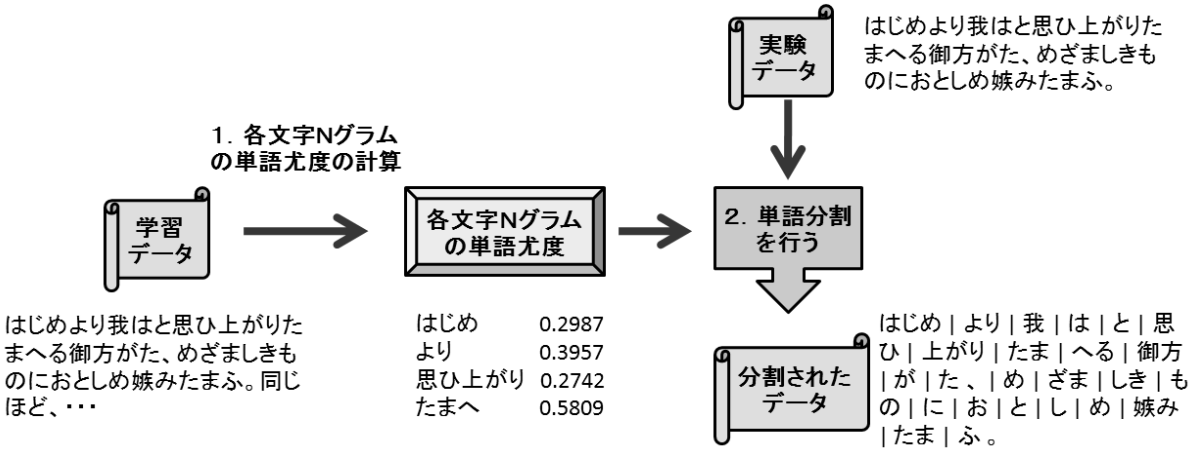


図 1：処理手順の概要

3. 文字 N グラム

文字 N グラムとは，与えられた文字列から *n* 文字を連続して切り出して得られる部分文字列のことである．まず，与えられた文字列から先頭の *n* 文字を切り出し，最初の文字 N グラムを得る．次に，与えられた文字列から 1 文字ずらし，つまり 2 文字目から *n* 文字を切り出し，次の文字 N グラムを得る．以下同様の処理を行い，与えられた文字列の最後の文字に達するまで繰り返すことにより，文字 N グラムを獲得する．例えば，「祇園精舎の鐘の声」 という文字列を 3 グラム（トライグラム）に分割すると，以下のようになる．

・祇園精
・園精舎
・精舎の
・舎の鐘
・の鐘の
・鐘の声

文字 N グラムの利点は，単語の境界が明確でない言語でも文字列を分割できることである．したがって，日本語や中国語のように，単語の境界が明確でない言語に対して用いられることが多い．そこで我々は，最初に日本語の古文テキストを文字 N グラムに分割し，単語の候補として扱うこととする．

4. 提案手法

4.1 提案手法の概要

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．

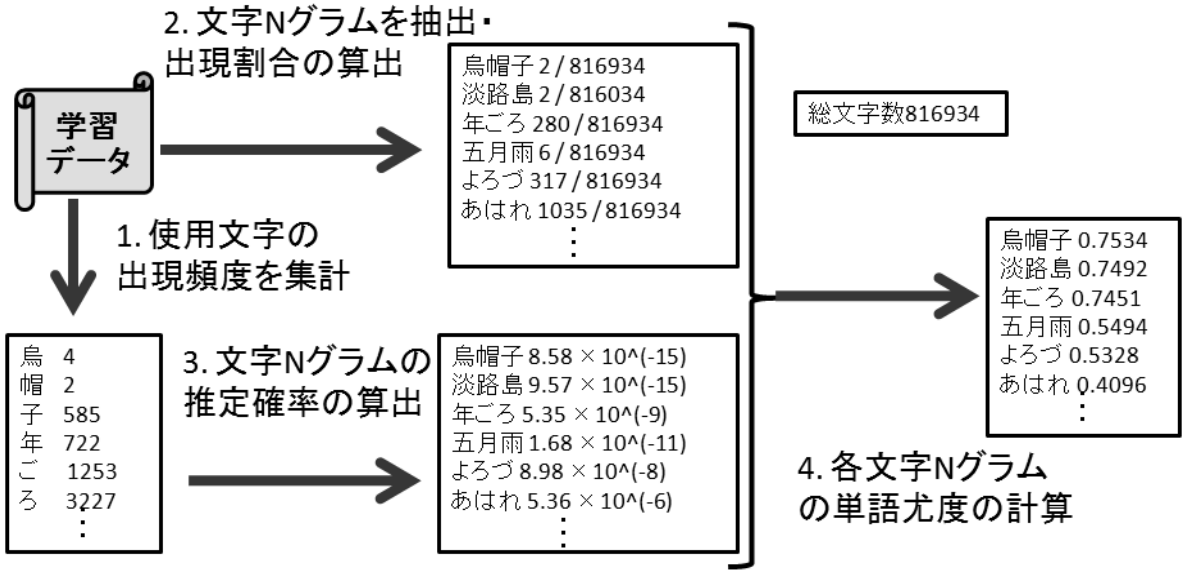
ここで， 単語である N グラムの出現頻度は，各文字の出現頻度から計算されるその N グラムの出現確率（以下「推定確率」と呼ぶ）よりもはるかに大きい頻度となるという仮定のもとで単語尤度 を定義する．「推定確率」は，文書中からランダムに *n* 文字抽出した際に，対象の N グラムとなる確率を意味する．この仮定より，単語尤度が高い N グラムを単語とみなすこととする．

4.2 単語尤度の計算方法

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 グラムの出現する割合との比をとることで単語尤度の計算式を定義する．しかし，単に比をとっただけでは N グラムの文字数が違うと計算結果の値が大きく変わり，異なる文字数の N グラム間で単語尤度の比較が行えない．そこで，異なる文字数間の比較を行うため，N単語共起傾向[4]の式を参考に式の拡張を行った．また，Normalized Pointwise Mutual Information (NPMI)[5]を参考に正規化を行った．

$$COT_{MIn}(w_1, w_2 \ldots w_n) = \frac{1}{n-1} \log_2 \frac{p(w_1, w_2 \ldots w_n)}{p(w_1)p(w_2)\cdots p(w_n)}$$

$p(w)$はコーパスにおける単語 w の存在する文書数の割合，$p(w_1,w_2 \ldots w_n)$は単語 $w_1,w_2 \ldots w_n$ の存在する文書数の割合である．



$$p(\text{烏})p(\text{帽})p(\text{子}) = \frac{4}{816934} \times \frac{2}{816934} \times \frac{585}{816934} = 8.58 \times 10\text{^}(-15)$$

図 2：単語尤度の計算の処理の流れ（3 グラム）

$$pmi(x;y) = \log \frac{p(x,y)}{p(x)p(y)}$$

$$npmi(x;y) = \frac{pmi(x;y)}{-\log p(x,y)}$$

N グラムの単語尤度 TL(*Term Likelihood*)を次の式で定義する．

$$TL(g) = \frac{\frac{1}{n-1}\log_2 \frac{p(g)}{p(x_1)p(x_2)\cdots p(x_n)}}{-\log_2 p(g)}$$

$p(g)$は N グラム g のコーパス内での出現頻度の割合，$p(x_n)$は N グラム g を構成する文字 x_n のコーパス内での出現頻度の割合を表している．

図 2 は 3 グラムの場合における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．

4.3 分割方法

求めた N グラムの単語尤度を用いて単語分割を行う方法を説明する．

1. 対象の文から先頭の 10 文字，なければそこまでを取り出す．
2. 取り出した中で一番大きい単語尤度の N グラムを見つける．
3. 取り出した文字列の中で 2 で見つけた N グラムの左側の文字列を全て取り出す．
   (ア) 2 で見つけた N グラムが先頭にあり，その直前が取り出せないとき，その N グラムを抽出する．
   (イ) 取り出した N グラムが 1 か 2 グラムの場合，その N グラムを抽出する．
   (ウ) 取り出した N グラムが 3 グラム以上の場合，2 に戻り処理を繰り返す．
4. 抽出した N グラムの次の文字から 10 文字取り出す．その後 2,3 を繰り返す．

この処理を，抽出していない文字がなくなるまで繰り返すことで文章の分割を行う．

図 3 は，求めた単語尤度を用いて単語分割を行う流れの例を示している．まず，網羅的に単語尤度を比較するために，対象の文の先頭から使用する N グラムの最大の文字数分を取り出す．この例では，今回の実験に合わせて 10 文字取り出している．次に取り出した中で一番大きい単語尤度の N グラム「右近」を見つける．それから，見つけた N グラムの左側の文字列全て「今日は」を取り出す．次にまた，取り出した中で一番大きい単語尤度の N グラム「今日」を見つける．「今日」は先頭にあるので，この N グラムを抽出する．最後に「は右近」からの 10 文字を取り出す．

この後，処理を文の最後まで続けることで分割を行う．

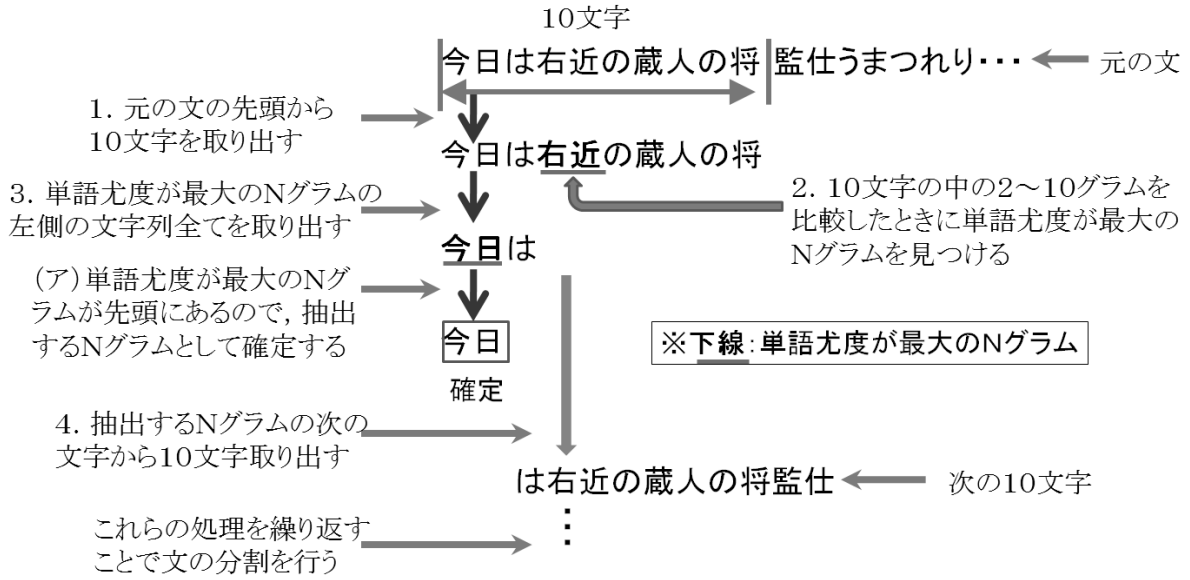


図 3：単語分割の流れの例

5. 実験

5.1 実験

前節で述べた提案手法により文章の単語分割を行う実験を行った．本手法では，1 グラムの値は計算を行うことができない．よって，使用する N グラムの下限は 2 となる．また，本手法では単語尤度の高い文字 N グラムを単語として取り出すため，使用する N グラム以上の文字数の単語数は検出することができない．しかし，日本語の古文において 10 文字を超える単語というのはほとんど存在しない．実際，今回評価に使用している「中古和文 UniDic」[1]による「源氏物語」[6]の解析では 10 文字を超える単語は検出されなかった．よって今回の実験では，2~10 グラムの N グラムを利用し実験を行うこととした．

本実験では「源氏物語」[6]を対象となる古典史料として用いた．「源氏物語」は全 54 巻あり，そこには全部で 816,934 文字記述されている．今回は，学習データと実験データを共に「源氏物語」の全 54 巻の同一のものを利用した．学習データより算出した各 N グラムの単語尤度は，学習データを分割する際に最も有効に使うことができる．よって，最終的には古文テキストの解析を目的とする当研究において，何よりもまず正しく単語を抽出することが重要であるため，学習データと実験データに同じデータを用いて最も良い結果を得ることは意味がある．まず，学習データより 2~10 グラムの単語尤度を計算し，次に，その単語尤度の値を用い実験データの単語分割を行う．

低頻度の文字 N グラムは単語尤度を求める際に悪影響を及ぼす可能性があるため，対象となる出現頻度の下限を 4 と設定している．悪影響とは，低頻度の N グラムが過大評価され単語として抽出されてしまうことである．また，単語尤度にも下限を設定している．今回の手法では 1 グラムに対して値が与えられないので，1 文字の単語に対するアプローチとして単語尤度の下限を 0.2 とし，それ以下の値の場合 1 文字で分割している．出現頻度の下限，単語尤度の下限共に，事前に行った予備実験の結果良い結果が得られた値を使用している．他に，空白，記号等は分割点であることが自明であるため，無条件で分割されるように設定している．しかし，評価の際には正解として数えないようにしている．

評価に際して関連研究で挙げた「中古和文 UniDic」[1]を使用している．「中古和文 UniDic」の解析は，源氏物語の一部を含む評価データに対して 95%以上の精度があるため，「中古和文 UniDic」の解析結果を評価用の正解データとして利用することが可能である．実験データを提案手法により分割を行った結果と，「中古和文 UniDic」による分割を行った結果を比較し，一致するものを正解とする．その正解数と，実験による分割数から適合率，「中古和文 UniDic」による分割数から再現率を算出し，最後に F 値を算出した．F 値とは，適合率と再現率の調和平均のことである．適合率 *precision*，再現率 *recall*，F 値 *F-measure* は次の式で算出する．

$$precision = \frac{R}{N}$$

$$recall = \frac{R}{C}$$

$$F-measure=\frac{2\times precision\times recall}{precison+recall}$$

R は正解した単語の数，N は「中古和文 UniDic」解析結果の単語の総数，C は実験による分割結果の N グラムの総数を表している．

これらは，単語の数だけでなく，単語境界の数からも算出している．「中古和文 UniDic」の解析結果の単語境界と，実験の分割結果の N グラムの境界を比較することで正解を判断し，それぞれの総数から F 値を算出している．さらに，単語の文字数別にも F 値の算出を行った．また，「中古和文 UniDic」による形態素解析の結果をもとに，品詞別の再現率を算出した．

5.2 実験結果

提案手法を用いた分割結果の一部を図 4 に示す． 図では，正しく分割された単語と境界，誤って分割された境界，本来分割されるべきだがされなかった境界を示している．

はじめ | **より** | **我** | **は** | **と** | 思ひ ▌上がり | たま▌へ⋮る | 御⋮方▌が▌た | 、 | め▌ざま▌しき | **もの** | **に** | お▌と▌し▌め | **嫉み** | たま▌ふ | 。 | **同じ** | **ほど** | 、 | そ▌れ | **より** | **下臈** | **の** | **更衣** | た▌ち | **は** | 、 | ま▌し▌て | やす▌か▌ら⋮ず | 。

文字Nグラム 正しく分割された単語
| 正しく分割された境界
▌ 間違って分割された境界
⋮ 本来分割されるべきだがされなかった境界

図 4：提案手法を用いた分割結果の一部

単語と単語境界の正解数から算出した適合率，再現率，F 値をそれぞれ表 1，表 2 に示す．また，単語の文字数別の適合率，再現率，F 値を表 3 に示す．品詞別の正解数，単語数，再現率を表 4 に示す．なお，本実験では，取り出した単語が何の品詞であるかの判定を行っていないので，品詞別の評価では適合率，F 値を算出することはできないため，これらの評価については行っていない．

表 1：単語ごとの評価結果

適合率	再現率	F 値
0.4550	0.5396	0.4937

表 2：単語境界ごとの評価結果

適合率	再現率	F 値
0.6668	0.8199	0.7355

表 3：文字数別の評価結果

文字数	適合率	再現率	F 値
1 文字	0.4558	0.6209	0.5257
2 文字	0.4611	0.6545	0.5411
3 文字	0.4553	0.2358	0.3107
4 文字	0.3868	0.1520	0.2182
5 文字	0.2769	0.0891	0.1348
6 文字	0.0239	0.0035	0.0062

表 4：品詞別の再現率

品詞	正解数	単語数	再現率
代名詞	3277	6668	0.4915
助詞	93842	130735	0.7178
接頭辞	3771	9723	0.3878
名詞	53341	86720	0.6151
副詞	12407	17689	0.7014
動詞	33403	92899	0.3596
助動詞	26668	57091	0.4671
形容詞	5004	23225	0.2155
接尾辞	2992	6977	0.4288
形状詞	1369	4806	0.2849
連体詞	3	13	0.2308
接続詞	18	805	0.0224
感動詞	58	320	0.1813
補助記号	12	28	0.4286
記号	3	6	0.5000
未知語	0	2	0.0000

5.3 考察

表 1 の単語の再現率をみると，正しい単語の 5 割以上が検出できていることがわかる．これより単語尤度の高い文字 N グラムが単語であるという仮説がある程度成り立っているといえる．また，表 2 の単語境界の再現率が 0.8199 と比較的高いことから，正しい単語境界については，ある程度検出できていることがわかる．しかし，それに比べ単語の評価結果の値や単語境界の適合率は，全体的に低くなっている．つまり，実際の単語より細かく分割されてしまっていると考えられる．この原因については，表 3 の考察と共に述べる．

表 3 の文字数別の評価結果をみると，長い単語になるほど再現率が低くなっていることがわかる．これは，今回の手法では 1 グラムに対しては単語尤度の値が与えられず，1 文字の単語を判別できないため，単語尤度の下限や分割方法により 1 文字の単語を判別しようとしたことで，全体的に文章が細かく分割されてしまって

いることが原因であると考えられる．つまり区切りの一方は正しいが，もう一方が不正確なため，単語全体を正確に抽出することができていないことが長い単語の再現率が低下する要因である．他にも長い単語の場合，その中にその単語より非常に頻出の N グラムを含む場合があり，その場合その非常に多く出ている N グラム方が優先して抽出され，単語の途中で不要に分割されてしまい，これも長い単語の再現率が低下する要因である．

また，1 文字の単語が 2 文字の単語より低い再現率になっている．これは，上で述べたように 1 グラムに対しては単語尤度の値が与えられず，うまく分割を行えないためである．そのため，1 文字の名詞と助詞が結合されてしまうパターンが多々見られる． 1 文字の単語がうまく分割できなかった例を表 5 に示す．

表 5：1 文字の単語の誤分割の例

正しい分割		分割結果
人|の|御|けはひ	→	人|の御|けはひ
先|の|世|に|も		先|の|世に|も
例|の|作法|に		例の|作法|に
前栽|の|露|は		前栽|の露|は

表 4 の品詞別の再現率をみると，名詞の正解率に比較して，動詞の再現率が低くなっていることがわかる．動詞の再現率が低いが，実際の古文テキスト解析では名詞の方が重要である場合が多いため，名詞の再現率が高いことは良い結果であるといえる．また，動詞の正解率が低くなっているのは，動詞の活用の影響だと考えられる．「たてまつる」という動詞を例に挙げると，「たてまつ」までは活用形によって変化しないので，「たてまつ」までの出現頻度が「たてまつる」に比べ非常に多い．結果，「たてまつる」は「たてまつ」と「る」に分割される傾向にある．動詞が誤って分割されてしまう例を表 6 に示す．

表 6：動詞の誤分割の例

正しい単語		分割結果
たまは	→	たま|は
たまひ		たま|ひ
たまふ		たま|ふ
たまへ		たま|へ

6. 現代語での実験

6.1 実験

前章で行った実験は，特定の時代の日本語古文を対象としており，またテキストの分量も限られている．そこで，より信頼性が高く大規模な評価実験として，現代日本語テキストに対し，学習データと重複しない実験データを用いて提案手法による単語分割を行う実験を行った．実験対象として，著作権が切れた文学作品を公開している「青空文庫」[1]中の，現代語で書かれている文書データを対象の文書として用いた．提案手法は基本的に言語に依存しないため，現代語でも古文の場合に近い精度が得られると考えられる．学習データとして，800 文書，約 870 万文字，実験データとして学習データと重複しない 89 文書，約 91 万文字を利用した．なお，振り仮名のルビ等は取り除いている．まず，学習データより 2~10 グラムの単語尤度を計算し，次に，その単語尤度の値を用い実験データの単語分割を行う．

その他の条件は，先の実験の条件と同じである．2~10 グラムを利用し実験を行い，対象となる出現頻度の下限を 4 と設定している．また，単語尤度の下限を 0.2 とし，それ以下の値の場合 1 文字で分割している．空白，記号等は無条件で分割されるように設定している．しかし，評価の際には正解として数えないようにしている．

評価に際して形態素解析器の MeCab を使用している．実験データを提案手法により分割を行った結果と，MeCab による分割を行った結果を比較し，一致するものを正解とする．先の実験の評価方法と同じように，適合率，再現率，F 値を算出する．

6.2 実験結果

提案手法を用いた分割結果の一部を図 5 に示す． 図では，正しく分割された単語と境界，誤って分割された境界，本来分割されるべきだがされなかった境界を示している．

終戦⋮後｜、｜**アメリカ**｜**が**｜**図書館**｜**界**｜**に**｜示▌し⋮た｜**関心**｜**は**｜ま▌こと▌に｜**深い**｜も▌の｜が⋮あ▌っ⋮た｜。｜**その**｜**一つ**｜**は**｜国立⋮国会▌図書館｜**の**｜**設立**｜で⋮あ▌り｜、｜**その**｜**二**｜**は**｜**図書館**｜**法**｜**の**｜**制定**｜で⋮あ▌り｜、｜**その**｜**三**｜**は**｜、｜**図書館**｜**学校**｜**の**｜ため⋮に｜**アメリカ**｜**の**｜**費用**｜**で**｜**アメリカ**｜**教師**｜**を**｜遣⋮わ⋮し▌た｜**こと**｜**で**｜**ある**｜。

文字Nグラム 正しく分割された単語
｜ 正しく分割された境界
▌ 間違って分割された境界
⋮ 本来分割されるべきだが
されなかった境界

図 5：提案手法を用いた分割結果の一部

[1] http://www.aozora.gr.jp/

単語と単語境界の正解数から算出した適合率，再現率，F 値をそれぞれ表 7，表 8 に示す．また，単語の文字数別の適合率，再現率，F 値を表 9 に示す．品詞別の正解数，単語数，再現率を表 9 に示す．

表 7：単語ごとの評価結果

適合率	再現率	F 値
0.5335	0.5819	0.5567

表 8：単語境界ごとの評価結果

適合率	再現率	F 値
0.7464	0.8241	0.7833

表 9：文字数別の評価結果

文字数	適合率	再現率	F 値
1 文字	0.5423	0.6480	0.5905
2 文字	0.5369	0.5739	0.5548
3 文字	0.3850	0.2744	0.3204
4 文字	0.5885	0.1754	0.2703
5 文字	0.1129	0.0300	0.0473
6 文字	0.5217	0.0738	0.1294

表 10：品詞別の再現率

品詞	正解数	単語数	再現率
名詞	119332	182327	0.6545
助詞	123314	171139	0.7205
動詞	24247	82265	0.2947
助動詞	26264	58516	0.4488
副詞	7152	16885	0.4236
形容詞	3569	9684	0.3685
連体詞	4163	7355	0.5660
接続詞	1393	4941	0.2819
接頭詞	4430	5401	0.8202
フィラー	291	405	0.7185
感動詞	382	1630	0.2344
記号	240	438	0.5479
その他	9	9	1.0000

6.3 考察

この現代語での実験の結果を，先の古典史料での実験結果と比較すると，表 7~表 10 のどの結果をみても同じような傾向の結果が出ていることがわかる．このことから，今回実験で利用した文書だけでなく，他の時代の文書でも同じような結果が出る可能性か高いことが示された．

7. あとがき

本論文では，日本語の古文テキスト解析のための単語分割手法の提案を行った．本手法では，単語分割のためにまず，古文のテキストから各文字 N グラムの単語尤度の計算を行う．次に，計算した各文字 N グラムの単語尤度の値を用いて古文のテキストの単語分割を行う．実験の結果，単語尤度の値が高い文字 N グラムが，正しい単語である可能性が高いことが示された．

今回の実験は，日本語の古文を対象に行った．しかしながら，我々の提案手法は言語によらないものであり，また辞書などの言語資源を一切必要としないことも特徴である．そして，日本語のような単語の境界が明示的でない言語に対しても利用することができる．それらのことから，古代中国の漢文など，他の言語に対して本手法を適用することも可能である．

我々の手法は，古文テキストの解析や情報抽出，辞書の構築，テキストマイニングだけでなく，テキストマイニングの結果の可視化などの人文科学の分野の様々なアプリケーションの基礎として利用することができる．

今後は，提案手法の改善を行い，精度を向上させる必要がある．精度の向上を目指して，1 文字の単語への対処方法や，途中で分割されてしまう単語へ対応，分割方法の改良を行っていく．1 文字の単語への対処方法として，1 グラムに何らかの値を与えられないか，単語の途中で分割されてしまう単語へ対応として，一度分割した前後の文字 N グラムを条件付けで再び結合できないかを検討している．

謝辞

本研究の一部は文部科学省グローバル COE プログラム「日本文化デジタル・ヒューマニティーズ拠点」，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)「言語・時代・文化横断型の情報アクセスに関する研究」(研究代表者:前田亮, 課題番号:21700271)，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)の支援を受けている．

参考文献

[1] 小木曽智信,小椋秀樹,田中牧郎,近藤明日子,伝康晴: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 人文科学とコンピュータ, 2010-CH-85,pp.1-8,2010.

[2] Gakuto Kurata, Shinsuke Mori, and Masafumi Nishimura. Unsupervised adaptation of a stochastic language model using a Japanese raw corpus. In Proceedings of the ICASSP2006, Vol.1, pp.1037-1040, Toulouse, France, May 2006.

[3] 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 自然言語処理,2009-NL-190,pp49,2009.

[4] 前田 亮, 吉川 正俊, 植村 俊亮 ：言語横断情報検索における Web 文書群による訳語曖昧性解消 ， 情報処理学会論文誌：データベース, Vol. 41, No. SIG 6 (TOD 7), pp. 12-21, 2000.
[5] Bouma Gerlof: Normalized (pointwise) mutual information in collocation extraction. In Proceedings of the Biennial GSCL Conference, 2009.
[6] 渋谷栄一:源氏物語の世界, http://www.genji-monogatari.net/