固有名詞の検出による古文並列コーパスを利用した時代横断対訳辞書の構築

木村 文則　　前田 亮
立命館大学 情報理工学部

近年，古い文書の電子テキスト化が進んでいるが，これらに対して現代語で行われているような自然言語処理技術を適用することは，十分な語彙量がある古語辞書などの言語資源が不足しているため現状では困難である．そこで本論文では，古文・現代語訳並列コーパスを用いた日本語の古語・現代語辞書の自動構築手法を提案する．古文・現代語訳並列コーパスとは，ある古文の文と，それに対する現代語訳の文が対になった文書群のことである．対となったコーパスを用いて，古文の文とその現代語訳の文に出現する単語を比較し，その出現傾向を解析することにより，古語と現代語との対応を獲得する．この際古語 N グラムを古語単語として扱うが，本手法では不要な N グラムを除去するために固有名詞を利用する．現代語訳の文に対し形態素解析を行い，現代語の固有名詞を抽出する．その文に対応する古文の文と抽出された固有名詞とのマッチングを行い，古文中の固有名詞を検出し，除去する．これにより，古語と現代語の対訳辞書の構築を行う．また，提案した手法による，現代語単語と古語 N グラムの出現傾向の類似度を算出する実験を行った．

Construction of Ancient-Modern Word Dictionary from Parallel Corpus of Ancient Writings and Their Translations in Modern Language Using Proper nouns Detection

Fuminori Kimura　　Akira Maeda

College of Information Science and Engineering,
Ritsumeikan University

Recently, an increasing number of ancient documents are being digitized in text form, but it is difficult to apply natural language processing techniques to these documents because the language resources for ancient languages, such as archaic dictionaries that have sufficient vocabularies, are scarce. In this paper, we propose a method for constructing an ancient-modern Japanese dictionary using parallel corpus of ancient writings and their translations in modern language. The parallel corpus consists of document pairs in the same language but in ancient and modern versions. From this corpus, we try to acquire equivalent pairs of archaic and modern word by analyzing the frequencies of word occurrences in a sentence in ancient language and its corresponding modern language translation. We deal with N-grams of archaic terms instead of original archaic terms. Our proposed method utilizes proper nouns in order to remove needless N-grams. We conduct morphological analyze for sentences of translations in modern language and extract proper nouns. And we conduct string match between extracted modern proper nouns and original archaic sentence in order to detect archaic proper nouns. Detected archaic proper nouns are removed from original archaic sentence. In this way, we construct an ancient-modern word dictionary. Besides, we conducted an experiment of calculating similarities of occurrence frequencies of archaic and modern words.

1．はじめに

ディジタル技術の進歩により，古典資料をディジタル化することにより保存が可能となった．ディジタル化が行われ始めた当初は，古典資料を保存することが目的で，その資料の画像データやテキストデータを残すことに主眼が置かれていた．

最近では徐々にではあるが，古典資料の本文やその現代語訳をテキストデータ化し，公開されるようになってきている．ディジタル化されたテキストはコンピュータによる処理が可能である．このことは，

これまでに培われてきた自然言語処理の技術が古典資料にも適用できる可能性があることを意味する．

日本語において自然言語処理を行う際には，形態素解析を行うことが多い．日本語の文章は，英語などのように単語と単語の区切りが明示的に表現されていないことから，単語に分割する必要があるためである．形態素解析ツールは，事前に作成された辞書を基に形態素解析を行う．しかし，古文に関しては形態素解析を行うのに必要な辞書がなく，利用できたとしても語彙数が十分ではないため，古文を対象とした形態素解析ツールは現在のところ公開されていない．それゆえ，現状では古典資料に対して自然言語処理を適用することは困難である．

現状の古文の言語資源不足に鑑み，本論文では古文の並列コーパスを用いた時代横断対訳辞書の自動構築を行う．古文の並列コーパスとは，ある古文の文章と，それに対する現代語訳の文章が対になった文書のことである．対となったコーパスを用いて，古文の文章とその現代語訳の文章に出現する単語を比較し，その出現傾向を解析することにより，古語と現代語との対応を獲得する．これにより，古語と現代語の対訳辞書，すなわち「時代横断対訳辞書」の構築を行う．以前の研究[1]では，明らかに対応関係のない現代語と古語の N グラムの組み合わせが抽出されていたことが問題であった．そこで本論文では本文中に出現する固有名詞を利用し，不必要な現代語と古語の N グラムの組み合わせを除去することにより，時代横断対訳辞書の構築手法の改善を図る．

2. 関連研究

現代語においては，二言語コーパスを利用してある二つの言語（例えば日本語と英語）間での訳語の対応推定を行う研究が行われている．訳語の対応推定手法は，大きく分けて，文対応がつけられたコーパス（並列コーパス）を用いる手法と，文の対応がつけられていないコーパスを用いる手法の二種類に分けられる．

文対応がつけられたコーパスを用いる手法では，訳語候補となる語の組に対して，共起頻度や分割表などを用いて統計的な相関を測定することにより，訳語の対応の推定を行う手法がよく知られている[2]．

文の対応がつけられていないコーパスを用いる手法では，一般に，訳語候補となる語の組に対して，何らかの方法により文脈の類似性を測定することにより，訳語候補の順位づけをおこない，訳語の対応の推定を行う[3]．

本論文は，現代語の二つの言語間ではなく，一つの言語の現代語と古語を対象として，訳語の対応の推定を行う．

現代語を対象とした場合，コーパスとして利用できる言語資源が豊富にあり，入手は容易である．それに対し，古文を対象としたコーパスは非常に乏しく，十分な量のコーパスを収集するのは困難である．しかしながら，著名な古典作品においては対訳が行われていることも多く，並列コーパスを入手することはそれほど困難ではない．

このような状況を考慮し，本論文では古文を対象とした並列コーパスを用いることにより，現代語と古語の訳語の対応の推定を行う．並列コーパスを用いることから，本論文の手法は，文対応がつけられたコーパスを用いる手法の範疇に属する．

3. 提案手法

本論文では，古文・現代語訳並列コーパスを用いて古語と現代語の対訳辞書の構築を行う手法を提案する．著名な古典作品では，現代語訳がなされていることも多い．さらに近年ではそれらのうちのいくつかは電子化され，公開されているものもある．

古文・現代語訳並列コーパスにおいては，古文の文とその翻訳である現代語の文の間の対応を取ることがある程度可能である．古文の文中に出現したある古文単語に対応する現代語の単語は，その古文の文と翻訳関係にある現代語の文において出現している可能性が高い．また，その逆も同様であるといえる．つまり，翻訳関係にある古文の文と現代語の文において共起頻度が高い古語と現代語は，対訳関係にある可能性が高い．

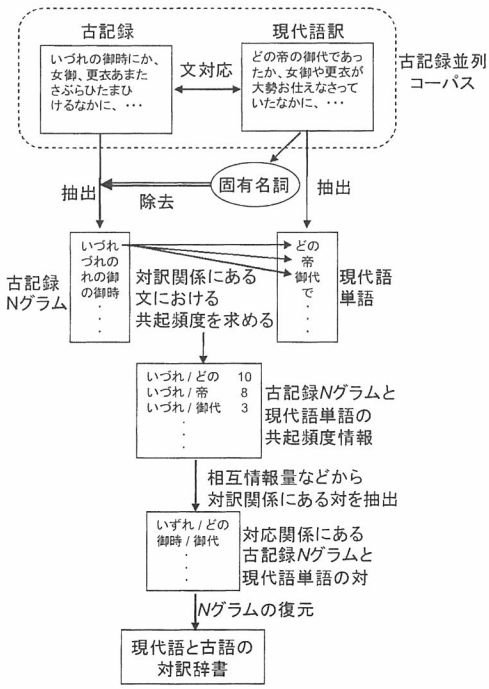

図1　時代横断対訳辞書構築の流れ

そこで本研究では，古文・現代語訳並列コーパスにおいて，古文とその現代語訳の間で出現する単語の統計情報から，翻訳関係にある文間での古語と現代語の単語の出現傾向の類似性を算出する．この出現傾向の類似性をもとに，対訳関係にある古語と現代語の単語の対応関係を導き出す．

本手法の処理の手順は以下のとおりである．

1. 並列コーパスからの単語抽出
2. 現代語単語と古語の共起頻度の算出
3. 現代語単語と古語の出現傾向の類似度の算出
4. 対応関係の認められる現代語単語と古語の抽出

図 1 は，提案手法による時代横断対訳辞書構築の流れを示している．

まず，現代語訳に対して形態素解析を行い，単語を抽出する．一方，現状では古文を対象とした形態素解析ツールが無いため，古文の文章を単語に切り分けることができない．それゆえ，古文は N グラムにより文章を切り分ける．古文の文章を N グラムにより文章を切り分ける前に，固有名詞で文章を分割し，その固有名詞を除去する．古語の固有名詞の検出は，対応する現代語訳の文章中に現れる固有名詞との文字列マッチングにより行う．その後，固有名詞で分割された古文の文章を N グラムに分割する．

次に，抽出された現代語の単語および古文の N グラムの共起頻度を算出する．本手法では，古文並列コーパスの文対応が取れている文単位で処理を行う．

古記録に出現するある N グラムに対して，その N グラムが出現する文に対応する現代語訳の文中に出現する現代語の単語を抽出する．こうして抽出された現代語の単語は，その N グラムと共起するとみなす．

上記で得られた共起頻度の統計情報から相互情報量を求める．その値が一定以上となる現代語の単語と古文の N グラムとの対を，古語と現代語の対訳関係があるとみなし，抽出する．

さらに，その現代語に対応付けられた古文の N グラムが単語であるかを調べる．文字数が N を超える単語の場合，N グラムで表現すると分割されてしまうため，分割された N グラムから単語に復元する必要があるからである．同じ現代語の単語と共起する古文の N グラムで，同士の統計情報を比較し解析することにより，N グラムから古語の単語を復元する．

上記の結果より現代語の単語と古語の単語との対応関係を導き，時代横断対訳辞書を構築する．

3.1 並列コーパスからの単語抽出

まず，現代語訳の文に対して形態素解析を行い，単語を抽出する．一方，古文の文に対してであるが，本来であれば現代語訳の文の場合と同様に単語を抽出することが望ましい．しかし，現状では古文を対象とした形態素解析ツールが無いため，古文の文を単語に切り分けることができない．それゆえ，古文は N グラムにより文を切り分け，これを単語として扱うこととする．

N グラムとは，文字列を単語単位ではなく一定の N 文字単位で分解したもののことである．まず，文の先頭から N 文字切り出す．次に，先頭から 1 文字ずらして N 文字切り出す．以下同様に 1 文字ずらしてから N 文字切り出すことを文の最後まで繰り返す．例えば，「祇園精舎の鐘の声」という文を 3 グラムで分解すると，以下のように分解される．

“祇園精”
“園精舎”
“精舎の”
“舎の鐘”
“の鐘の”
“鐘の声”

単語単位で分割するのに比べ，N グラムは重複が多いが，単語の境界が明確でなくても文字列を分割できるため，日本語のように単語の境界が明確でない言語の文に対して用いられることが多い．本論文の対象である古文の文も単語の境界が明確でないため，N グラムに文を分割し，これを古文の単語として扱うこととする．この N グラムのことを，これ以降「古語 N グラム」と呼ぶ．

本手法では，古文の文章を N グラムに分割する前に，固有名詞で文章を分割し，その固有名詞を除去する．古語の固有名詞の検出は，対応する現代語訳の文章中に現れる固有名詞との文字列マッチングにより行う．その後，固有名詞で分割された古文の文章を N グラムに分割する．これにより，不必要な N グラムが生成されなくなるため，現代語と古語 N グラムとのマッチング精度の低下を抑えることができる．

3.2 現代語単語と古語の共起頻度の算出

3.1 節において抽出された現代語単語および古語 N グラムの共起頻度を算出する．本手法では，古文並列コーパスの文対応が取れている文単位で処理を行う．つまり，翻訳関係にある古文の文と現代語の文において同時に出現する古語と現代語の組み合わせを，共起しているとみなす．

古文に出現するある古語 N グラムに対して，その N グラムが出現する文に対応する現代語訳の文中に出現する現代語の単語を抽出する．こうして抽出された現代語の単語は，その N グラムと共起するとみなす．

ある古語 N グラムとある現代語単語の組み合わせの共起が，対象となる古文並列コーパスの全文中で何度起こるか数え上げ，その回数をその古語 N グラムと現代語単語の組み合わせの共起頻度とする．

3.3 現代語単語と古語の出現傾向の類似度の算出

3.1 節において抽出された現代語単語および古語 N グラムから得られる出現頻度や，3.2 節で得られた古

語 N グラムと現代語単語の組み合わせの共起頻度の統計情報から，現代語単語と古語 N グラムの出現傾向の類似度を算出する．

日本語と英語のように異なった言語間において，並列コーパスから対訳関係にある単語同士の出現傾向の類似度を算出する手法として，相互情報量やDice係数を用いる手法が提案されている [3]．本論文においても，相互情報量を用いて現代語単語と古語 N グラムの出現傾向の類似度を算出する．

・ 相互情報量

ある現代語単語 t と古語 N グラム g における相互情報量 $MI(t, g)$ は，以下の式から算出される．

$$MI(t,g) = \log \frac{P(t,g)}{P(t)P(g)} = \log \frac{\frac{f(t,g)}{N_{pair}}}{\frac{f(t)}{N_c}\frac{f(g)}{N_a}}$$

確率 P(t, g) は，古語 N グラム g が古文の文中に出現し，なおかつその文と翻訳関係にある現代語の文中に現代語の単語 t が出現する確率を表す．また，確率 P(g) は，古語 N グラム g が古文の文中に出現する確率を表す．3 つの確率 P(t, g) ，P(t), P(g) はいずれも，現代語の単語 t と古語 N グラム g の間の共起頻度および g の出現頻度から算出することができる．$f(t, g)$ は現代語の単語 t と古語 N グラム g の間の共起頻度，$f(t)$, $f(g)$はそれぞれ現代語の単語 t と古語 N グラム g の出現頻度，$Npair$ は現代語の単語と古語 N グラムの組合せの総出現数，Nc, Na はそれぞれ現代語単語および古語 N グラムの総単語数を表す．こうして求められた相互情報量の値が大きい古語と現代語の組み合わせほど，古文並列コーパスにおける出現傾向が類似しているといえる．

上記において求めた相互情報量の値を現代語単語と古語 N グラムの出現傾向の類似度とする．この類似度が高い古語 N グラムと現代語の組み合わせほど，出現傾向が類似しているといえる．すなわち，対訳関係にある古語 N グラムと現代語の組み合わせである可能性が高い．その類似度が一定以上となる現代語単語と古語 N グラムとの対を，古語と現代語の対訳関係があるとみなし，抽出する．

3.4 対応関係の認められる現代語単語と古語の抽出

3.3 節において，古語と現代語の対訳関係がある可能性の高い現代語単語と古語 N グラムとの組み合わせを抽出したが，このとき抽出された N グラムは必ずしも古語の単語となっているとは限らない．機械的に N 文字に分割しているため，その N グラムはある古文単語の一部分である可能性がある．また，前後に別の古語の一部が結合されている可能性もある．そのため，N グラムから古語の単語を復元することが必要となる場合も起こる．

対訳関係がある可能性の高い組み合わせの現代語において，その現代語の単語と共起する別の N グラムと，出現頻度やその現代語との共起頻度などの統計情報を比較し解析することにより，N グラムから古語の単語を復元する．こうして復元された古文単語と抽出された組み合わせの現代語の単語が，互いに対訳関係になっているとみなす．

上記の結果より現代語の単語と古語の単語との対応関係を導き，古語・現代語辞書を構築する．

4. 実験

3 章において提案した手法により，古文・現代語訳並列コーパスから現代語単語および古語 N グラムの組み合わせを抽出し，その現代語単語と古語 N グラムの出現傾向の類似度を算出する実験を行った．

古文・現代語訳並列コーパスとして，源氏物語の定家本系『源氏物語』（青表紙本）本文およびその現代語訳を用いた [4]．本実験では，第 1 巻から第 20 巻までの合計 20 巻を実験対象とした．

古文・現代語訳並列コーパスの現代語訳からの現代語単語の抽出の際，文を単語に切り分けるために形態素解析ツールである ChaSen を用いた．本実験では，品詞の種類による選別は行わず，得られた現代語単語は全て抽出された単語として使用した．その結果，抽出された現代語単語は 108,361 単語であった．

また，古文・現代語訳並列コーパスから古語 N グラムの抽出は，3 グラム（trigram）の場合について行った．その結果，抽出された N グラム総数は，197,357 個であった．

現代語単語および古語 N グラムの組み合わせを，古文並列コーパスの文対応が取れている文単位で作成した後，これらの組み合わせの共起頻度を求めた．共起頻度も，3 グラムの場合において求めた．その結果，抽出された現代語単語および古語 N グラムの組み合わせの総数は，4,631,886 組であった．

現代語単語と古語 N グラムの出現傾向の類似度の算出は，相互情報量を用いて求めた．出現傾向の類似度の算出においても，共起頻度の場合と同様に，3 グラムの場合について算出を行った．ただし，相互情報量では低頻度語が過大評価される傾向があることが知られており[5]，以前の研究[1]においてもその悪影響が生じている．そのため，本実験においては，現代語の単語および古語 N グラムの出現頻度が 5 未満である組合せは，出現傾向の類似度算出の対象から除外した．

5. 考察

表 1，**表 2** は，4 章の実験結果における，現代語「夕顔」に対する相互情報量の大きい古語の 3 グラムの上位 10 件を示したものである．表 1 は固有名詞

を除去しない場合，表 2 は提案手法である固有名詞を除去した場合の結果である．なお，源氏物語において「夕顔」は登場人物の一人であるため，本来であれば固有名詞として扱われるべきであるが，本実験において現代語訳を形態素解析した結果では，一般名詞と判定された．

表 1 現代語「夕顔」に対する相互情報量の大きい古語 3 グラムの上位 10 語（固有名詞を除去しない場合）

現代語	Nグラム	相互情報量
夕顔	りて見	17.3998
夕顔	ましさ	17.3998
夕顔	の契り	17.3998
夕顔	とけぬ	17.3998
夕顔	たの御	17.3998
夕顔	く思ほ	17.3998
夕顔	くうち	17.0915
夕顔	気色ば	16.8244
夕顔	ゆるを	16.3781
夕顔	て参れ	16.3781

表 2 現代語「夕顔」に対する相互情報量の大きい古語 3 グラムの上位 10 語（固有名詞を除去した場合）

現代語	Nグラム	相互情報量
夕顔	**夕顔の**	**28.5240**
夕顔	の内の	28.5240
夕顔	し忘れ	28.5240
夕顔	しあは	28.5240
夕顔	知りは	28.0615
夕顔	思ほえ	28.0615
夕顔	思し忘	28.0615
夕顔	限りの	28.0615
夕顔	もかし	28.0615
夕顔	の御い	28.0615

表 1，表 2 において，「夕顔」という文字列を含んでいる古語 N グラムが，現代語単語「夕顔」に対して適切な組合せである．表 1 では，「夕顔」という文字列を含んでいる古語 N グラムは上位には現れておらず，適切な組み合わせを抽出することができていない．それに対し表 2 では，最上位に「夕顔」という文字列を含んでいる古語 N グラムが現れており，適切な組み合わせを抽出することが可能であることから，本手法による現代語と古語 N グラムとのマッチング精度の改善が認められる．

しかし，本手法において現代語と古語 N グラムとのマッチング精度がまだ十分に改善されていない場合もある．**表 3** は，4 章の実験結果における，固有名詞を除去した場合の現代語「某」に対する相互情報量の大きい古語の 3 グラムの上位 10 件を示したものである．現代語「某」に対応する古語として「なにがし」があげられることから，「なにがし」の部分文字列を含む古語 N グラムが適切な組合せであるといえる．表 3 の結果では，上位 10 件には適切な古語 N グラムが現れていない．適切な古語 N グラムである「なにが」，「にがし」が出現するのは上位 30 件目であり，これらの古語 N グラムは抽出されない．

表 3 現代語「某」に対する相互情報量の大きい古語 3 グラムの上位 10 語（固有名詞を除去した場合）

現代語	Nグラム	相互情報量
某	世にか	28.8117
某	人はべ	28.8117
某	るしな	28.8117
某	やる方	28.8117
某	にわづ	28.8117
某	にせむ	28.8117
某	れかか	28.3492
某	だ今の	28.3492
某	いつの	28.3492
某	をも見	27.9486
	（中略）	
某	で立ち	26.0628
某	**にがし**	**25.8692**
某	**なにが**	**25.8692**
某	立ちて	25.5158

今回うまく抽出できなかった「なにが」，「にがし」の二つの古語 N グラムであるが，これらを連結すれば「なにがし」という古語を復元できる．また，これらの古語 N グラムの相互情報量は同じ値となっていることから，「なにがし」という古語の出現傾向が「なにが」，「にがし」の二つの古語 N グラムで表現できていると考えられる．よって，不必要な

古語 N グラムをさらに除去することができれば，このような適切な古語 N グラムを抽出し，古語を復元することが可能となる．

今回の実験において，本手法が現代語と古語 N グラムとのマッチング精度の改善を十分に行えない場合がある原因として，除去された固有名詞が少なかったことがあげられる．これは，本来は固有名詞であるはずの単語が，形態素解析を行った結果，一般名詞と判定されていることが多いためである．表 1，表 2 で取り上げた「夕顔」がその例である．この問題を改善するためには，一般名詞も固有名詞と同様の処理を行うことを検討する必要がある．

また，源氏物語において「桐壺」という登場人物がいるが，この固有名詞は形態素解析の結果，「桐」と「壺」という別々の単語として分割されてしまっている．しかも，いずれも一般名詞として解析されている．このような単語をもとの一単語として処理できるようにすれば，さらなる本手法の精度の改善ができると考えられる．

6. おわりに

我々は古文の並列コーパスを用いることにより，現代語の単語と古語の単語との対応関係を導き，時代横断対訳辞書を自動構築する手法の提案を行った．本論文では，固有名詞を検出することにより本手法の改善を図った．

古文書や古記録に対して自然言語処理技術を適用するには，古語の対訳辞書が必要不可欠である．しかし，現状では古語の言語資源は十分であるとはいえない．本手法は，古語の言語資源を自動的に構築する手法であり，このような状況を改善することに貢献できると考えている．

時代横断対訳辞書が充実することにより，古語の形態素解析などの応用も可能になると考えている．一般に，現代語の形態素解析を行うには現代語の辞書を必要とする．古語の形態素解析でも同様である．

このような技術が実現していくと，ゆくゆくは古文書や古記録そのものを解析することが可能となると考えている．その結果，古文書や古記録に関する研究にこれらの技術が貢献できるようになると思われる．さらには，古文に対する教育への応用なども考えられる．

今後の課題としては，形態素解析により一般名詞と判定されてしまう固有名詞に対する処理や，分割されてしまう単語に対する処理を改善することにより，本手法の精度のさらなる向上を図ることである．また，本論文において提案した手法を用いて実際に時代横断対訳辞書を構築する実験を行うことである．

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) の支援を受けている．

参考文献

[1] 木村 文則, 前田 亮: 古文・現代語訳並列コーパスによる古語・現代語辞書の構築, 人文科学とコンピュータシンポジウム論文集, pp. 119-124, 2008.

[2] Kitamura, M. and Matsumoto, Y.: Automatic Extraction of Word Sequence Correspondences in Parallel Corpora , *In Proceedings of the 4th Workshop on Very Large Corpora*, pp.79-87, 1996.

[3] Tanaka, T.: Measuring the Similarity between Compound Nouns in Different Language Using Non-Parallel Corpora, In *Proceedings of the 19th COLING*, pp.981-987, 2002.

[4] 渋谷栄一: 源氏物語の世界 http://www.sainet.or.jp/~eshibuya/

[5] 久光 徹, 丹羽 芳樹: 統計量とルールを組み合わせて有用な括弧表現を抽出する手法, 情報処理学会研究報告, NL-122-17, pp. 113-118, 1997.