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

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

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. 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

検索エンジンなどが実用化されたことにより，情報検索の技術が身近に利用できるようになった．一方で，様々な機関で古文書や古記録を含む古い文書（以降これらを総称して「古文」と呼ぶ）の電子テキスト化が進んでおり，検索可能な古文は今後ますます増加すると考えられる．しかし，現在の検索システムでは古文の検索は，古文の知識が乏しい利用者には容易ではない．なぜなら，問合せ語を古語で入力しなければならないからである．現状の検索システムで古文を対象に検索を行うためには，捜したい事柄に対する古語の表現を利用者自身が知っていなければならない．これは古文に精通していない利用者にとっては負担となる[1].

このことは，我々がこれまで研究を行ってきた言語横断情報検索[2]にも共通の問題である．言語横断情報検索とは，検索したい文書で使用されている言語とは別の言語の問合せを用いて検索を行う技術である（例えば，日本語の問合せを用いて英語の文献を検索するなど）．言語横断情報検索において，問合せに現代語を用い，検索対象を古文に置き換えれば，古文に対する検索に相当する．

言語横断情報検索では，入力された問合せ語をシステムが適切な言語に翻訳し，検索を行う．この方法は，古文を対象に検索を行う場合にも適用できると考えられる．これにより，利用者は時代の違いを意識することなく古文に対して検索を行うことが可能となる．古文を対象にした検索は「言語横断」というよりは「時代横断」検索であるといえる．以下では，古文を対象にした検索のことを，「時代横断型検索」と呼ぶ．

このような時代横断型検索を行うためには，現代語の問合せを古語に翻訳するための言語資源が必要である．しかし，現状では古語を対象とした対訳辞書さえ十分に整備されているとはいえず，利用できたとしても語彙数が十分ではない．

現状の古文の言語資源不足を鑑み，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の提案を行う．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．これにより，古語と現代語の対訳辞書の構築を目指す．また，提案した手法による，現代語単語と N グラムの出現傾向の類似度を算出する実験を行った．

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[3]．

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[4]．

本論文では，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．

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，提案手法による古語・現代語辞書構築の流れを示している．

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．

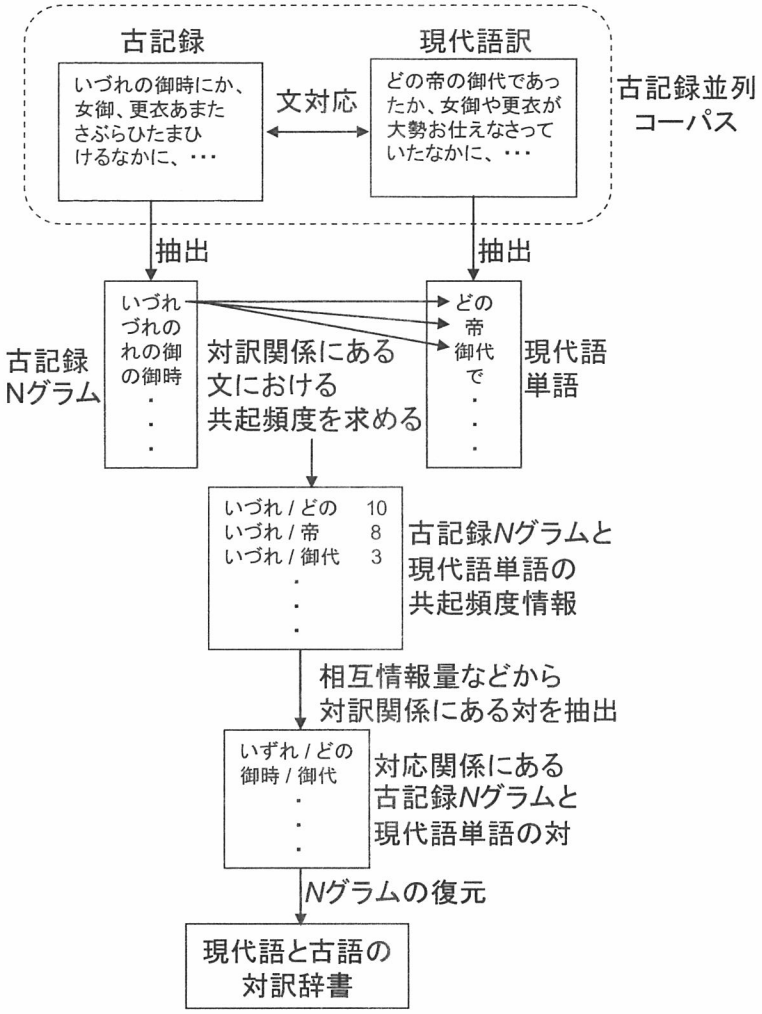

図 1 古語・現代語辞書構築の流れ

“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．

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，相互情報量および Dice 係数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．

$$MI(t,g) = \log \frac{P(t \mid g)}{P(g)}$$

確率 $P(t \mid g)$ は，古語 N グラム g が古文の文中に出現したときに，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．2 つの確率 $P(t \mid g)$，$P(g)$ 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．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．

・ Dice 係数

ある現代語単語 t と古語 N グラム g における Dice 係数 $sim(t, g)$ は，以下の式から算出される．

$$sim(t,g) = \frac{2f(t,g)}{f(t)+f(g)}$$

$f(t, g)$ は，現代語の単語 t と古語 N グラム g の間の共起頻度を表す．また，$f(t)$ および $f(g)$ はそれぞれ，t, g の出現頻度を表す．こうして求められた Dice 係数は，相互情報量の場合と同様に，そ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または Dice 係数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. 予備実験

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 [5]．本実験では，第 1 巻から第 10 巻までの合計 10 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，2 グラム（bigram），3 グラム（trigram），4 グラム，5 グラムの 4 とおり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，各グラム数で，106,956 個，95,755 個，85,225 個，75,155 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，2 グラムから 5 グラムの 4 とおり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，各グラム数で，1,058,690 組（2 グラム），1,431,018 組（3 グラム），1,516,490 組（4 グラム），1,454,701 組（5 グラム）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場合と，Dice 係数を用いて行った場合の両方について行った．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，グラム数別に算出を行った．

5. 考察

表 1 は，4 章の実験結果における，現代語「中宮」に対する相互情報量の大きい古語の 2 グラムの上位 10 件を示したものである．表 1 に示された全ての 2 グラムの相互情報量の値は 20.6884 となっている．このように相互情報量の値が同じ N グラムが多数上位に出現しているのは，共起頻度が 1 である現代語単語と古語 N グラムの組み合わせが多数存在していることが原因である．現代語単語と古語 2 グラムとの組み合わせ全 1,058,690 組のうち，773,589 組で共起頻度が 1 である．現代語「中宮」との組み合わせにおいても，最初に共起頻度が 1 でない組み合わせが出現するのは上位から 67 番目であった．

表 1 現代語「中宮」に対する相互情報量の大きい古語 2 グラムの上位 10 語

現代語	Nグラム	相互情報量
中宮	茂の	20.6884
中宮	法延	20.6884
中宮	父親	20.6884
中宮	孫王	20.6884
中宮	裳着	20.6884
中宮	召の	20.6884
中宮	重な	20.6884
中宮	子や	20.6884
中宮	御用	20.6884
中宮	御裳	20.6884

本手法では，共起頻度を算出する際に抽出する現代語単語と古語 N グラムの組み合わせは，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ため，対訳関係がない現代語単語と古語 N グラムの組み合わせも抽出される．本来であれば，このような組み合わせは共起頻度が低いため相互情報量の値も低くなることが予想される．しかし，今回の実験結果では，このような組み合わせが，共起頻度が 1 の場合でも上位になっているものも多い．これは，今回実験対象として用いた古文・現代語訳並列コーパスの量が不十分であったことが原因であると考えられる．

そこで，**表 2** では，共起頻度が 1 である組み合わせを除外した場合の，現代語「中宮」に対する相互情報量の大きい古語の 2 グラムの上位 10 語を示す．現代語「中宮」に対して，「中宮」という古語の 2 グラムは，明らかに対訳としての対応関係が認められる．

この「中宮」という古語の 2 グラムは，表 2 において，上位 2 番目に位置している．この結果より，現代語単語と古語 N グラムの相互情報量から，古文並列コーパス中での出現傾向の類似している組み合わせを抽出し，それをもとに対訳関係にある組み合わせを抽出することができる可能性があることを示している．

上記の傾向は，現代語単語と古語 N グラムの組み合わせの古文並列コーパス中での出現傾向の類似度に Dice 係数を用いた場合においても同様であった．

表 2 共起頻度が 2 以上である組み合わせにおける，現代語「中宮」に対する相互情報量の大きい古語 2 グラムの上位 10 語

現代語	Nグラム	相互情報量
中宮	ひい	17.7981
中宮	中宮	17.2227
中宮	女御	16.6586
中宮	賜は	16.5858
中宮	の女	16.4118
中宮	と忍	16.4118
中宮	はね	15.9655
中宮	みき	15.3901
中宮	せず	15.3901
中宮	宮に	15.1995

今回の実験においては，実験対象として用いた古文・現代語訳並列コーパスの量の不足からくる，共起頻度が 1 である現代語単語と古語 N グラムの組み合わせによる悪影響が最も影響していたといえる．この点に関しては，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[6]，これに対する対処としては，例えば出現頻度に閾値を設ける，Dice 係数に対して共起頻度を考慮した「改良 Dice 係数」[3]などを用いるなどが考えられる．

また，対訳関係にない現代語単語と古語の N グラムの組み合わせを多く抽出していることも問題の一つであると考えられる．本手法では，共起頻度を算

出する際に抽出する現代語単語と古語 N グラムの組み合わせは，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っているが，現在のところ，その文から得られる全ての現代語単語と全ての古語 N グラムとの組み合わせを抽出している．そのため，対訳関係にない現代語単語と古語 N グラムの組み合わせが大量に抽出される結果となっている．そのため，現代語単語と古語 N グラムとの組み合わせの抽出手法を改良し，対訳関係にない現代語単語と古語 N グラムの組み合わせが抽出される数を減らす必要がある．

例えば，既存の古語辞書を活用することが考えられる．既存の古語辞書から，明らかに対訳関係にある現代語単語と古語の組み合わせが幾つかでも特定できれば，その現代語単語と古語に関する組み合わせは抽出しなくてもよい．特に，古文の文の途中に特定できる古語があった場合，その箇所で文を分割できるため，生成される N グラムの数を減らすことが可能となる．これにより対訳関係にない現代語単語と古語 N グラムの組み合わせを減らすことができる．それにより，相互情報量または Dice 係数から算出した，古文並列コーパス中での出現傾向の類似度の精度が向上し，それをもとに対訳関係にある組み合わせを抽出する処理の精度も向上することが期待される．

6. おわりに

本論文では，古文・現代語訳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，古語・現代語辞書を自動構築する手法の提案を行った．また，提案手法により，古文・現代語訳並列コーパスから現代語の単語および古語の 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．

古文を検索するために，現代語を用いて検索する「時代横断型検索システム」を実現するには，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．

古語・現代語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．また，時代横断型検索などにも利用されることにより，古文に精通していない一般の利用者が古文に接することが容易になる．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，現代語単語と古語 N グラムとの組み合わせの抽出手法を改良し，対訳関係にない現代語単語と古語 N グラムの組み合わせが抽出される数を減らすことが挙げられる．また，本論文において提案した手法を用いて実際に古語・現代語辞書を構築する実験を行うことも，課題として挙げられる．

参考文献

[1] 木村 文則, 小牟礼 雅之, 前田 亮, 佐古 愛己, 杉橋 隆夫: 古典史料データベース検索システムの提案, 情報処理学会研究報告, 2008-CH-78, pp. 45-52, 2008.

[2] 木村 文則, 前田 亮, 波多野 賢治, 宮崎 純, 植村 俊亮: Web ディレクトリの階層構造を利用した問合せの分野推定に基づいた言語横断情報検索. 情報処理学会論文誌：データベース, Vol. 49, No. SIG 7 (TOD 37), pp. 59-71, 2008.

[3] 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.

[4] 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.

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

[6] 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.