日本語歴史コーパスの現代語辞書における未知語義判定システム

著者	田邊 絢，古宮 嘉那子，浅原 正幸，佐々木 稔，新納 浩幸
雑誌名	言語資源活用ワークショップ発表論文集
巻	3
ページ	112-117
発行年	2018
URL	http://doi.org/10.15084/00001643

日本語歴史コーパスの現代語辞書における未知語義判定システム

田邊 絢（茨城大学大学院理工学研究科）
古宮 嘉那子（茨城大学工学部情報工学科）
浅原 正幸（国立国語研究所コーパス開発センター）
佐々木 稔（茨城大学工学部情報工学科）
新納 浩幸（茨城大学工学部情報工学科）

Detecting Unknown Word Senses in Contemporary Japanese Dictionary from Corpus of Historical Japanese

Aya Tanabe (Ibaraki University)
Kanako Komiya (Ibaraki University)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Minoru Sasaki (Ibaraki University)
Hiroyuki Shinnou (Ibaraki University)

要旨

日本語歴史コーパス中の単語には、現代語と同様の意味で扱われている単語と、古語特有の意味を持つ単語がある。本研究では、この現代語にはない古語特有の単語の語義（言葉の意味）を未知語義と定義して、日本語歴史コーパス中から、未知語義を検出するシステムの提案を行う。具体的には、日本語歴史コーパス中の単語を、(1)現代の分類語彙表でその単語の分類番号として登録されている語義をもつ語、(2)現代の分類語彙表にある語義をもつが、現在その語義は、その言葉の語義として分類語彙表は登録されていない語、(3)その語義の定義が現代の分類語彙表にないため、分類番号が振られていない語、の 3 種類にクラス分けする。実験では、各単語について、出現書字形や見出しなどの 8 要素を基本素性として用いた。また、別の日本語歴史コーパスから word2vec を用いて、3 種類の単語の分散表現のベクトル(50 次元、100 次元、200 次元)を作成し、素性として加えた。それぞれ SVM を用いて正解率を比較したところ、日本語歴史コーパス中の未知語義の検出において、単語の分散表現のベクトルが正解率を向上させることが分かった。

１．はじめに

語義曖昧性解消をめぐるアプローチとしては、確率的な言語モデルに基づき、対象単語の前後にある単語の品詞や、各単語同士の共起関係などを特徴として用いて、コーパスから機械学習を行うなどの様々な試みがなされている。古文コーパスでの語義曖昧性解消を行う際に、現代文コーパスでの語義曖昧性解消のタスクをそのまま適用させようとすると、現代語と古語とでの単語の語義の違いから、現代語の分類で分類した語義には当てはまらず、新たに別の正しい語義を付与する必要のある古語の語義が存在する。このような現代文での分類においては未知となる古文の語義を、本研究において未知語義と呼ぶ。

現代文の語義曖昧性解消タスクに関する研究として、(Suzuki et al., (2018))がある。これは、コーパス中の全単語に対して、分類語彙表の類義語と分散表現を利用することで、対象単語とその類義語の周辺単語同士の類似性から語義を予測する手法を取っている。また、(遊佐ら, (2017))では、語義曖昧性解消タスクにおいて分類語彙表を用いることの有効性を示している。また、未知語の検出について、(新納ら, (2012))がある。これは、外れ値検出法を用いて、対象単語の語義が新語義となっている用例を検出する手法を提案している。

古文の語義分類に関して、(宮島ら, (2014))は古文作品における各単語の出現頻度に加えて、すべての単語に国立国語研究所編『分類語彙表』の分類番号を追記し、古語の分類語彙表としてまとめている。また、(小木曽, (2011))の古文用形態素解析辞書の開発に関する研究がある。これは、見出し語に短単位を採用することで、現代語と古語の仮名遣いの違いを考慮し、地の文と会話文とで別に辞書生成することで古文用の形態素解析辞書の解析精度が向上することを示している。

さらに、本研究に関連する研究として、、通時コーパスの構築に関する研究がある。しかし、古文と現代文の違いや、古文同士でも書かれた時代によって構文や語義が大きく異なることなどから、通時コーパスの構築に関しては解決すべき課題が多くある。この通時コーパスの構築に通ずる研究の一部として、まず歴史コーパスにおける語義曖昧性解消を行うために、現代文の語義分類に当てはまらない古文の未知語義を検出し、その種類を分類することを本研究の目的としている。

２． 歴史コーパス

コーパスは、自然言語処理の研究に用いるため、テキストや発話などの文章を構造化し、言語的な情報（品詞、統語構造など）を付与して、大規模に集積しデータベース化した言語資料である。特に古文の文章に対して構造化したものを、歴史コーパスと呼ぶ。また、歴史コーパスにおいては、データ化する過程で文字表記の関係で外字処理、テキストの校訂など表記の置き換えがなされている。歴史コーパスは各語に関して、出現書字形（orthToken）、出現発音形（pronToken）、語彙素読み（reading）、語彙素（lemma）、原文文字列（originalText）、品詞（pos）、古語品詞（sysCType）、活用形（cForm）、語彙素番号（lemmaID）、分類番号（wlsp）の要素からなっている。特に分類番号に関して、ピリオドのあるものは現代語の語義、ピリオドのないものは古語の語義、と区別がされている。

本研究で用いる歴史コーパス中の、現代語にある語義に分類されている語、現代語では別の語が持っている語義に分類されている語、語義分類されていない語（本研究においての未知語義を持つ語）の割合を以下の図１に示す。

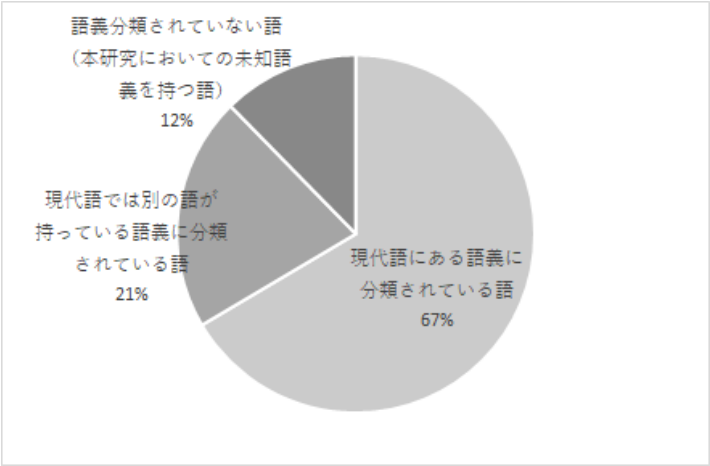

図１　歴史コーパス中の各分類語の割合

３．分散表現を利用した歴史コーパスにおける未知語義判定システム

３.１　原理

歴史コーパス中の語は、語義分類に関して、現代語にある語義に分類されている語、現代語では別の語が持っている語義に分類されている語、語義分類されていない語（本研究においての未知語義を持つ語）の 3 種類に分けられる。本研究ではこれらを分類した。、歴史コ

ーパス中の各語についている、形態素、品詞などの情報を、本実験のための素性として用いて学習データを生成した。また、歴史コーパスの情報に加えて、その語が単義語であるか多義語であるかの情報を素性として用いた。さらに、歴史コーパス中の各語について、別の歴史コーパスを用いて word2vec（Mikolov et al., 2013a, b, c）で生成した各単語の分散表現のベクトルを素性として用いた。生成した各学習データは、LIBLINEAR を用いて 5 分割の交差検定で学習を行う。

３.２　基本素性

本研究の学習データを生成する上での基本素性として、歴史コーパス中の各語について、出現書字形、出現発音形、語彙素などの 9 種類の要素を用いる。これらの要素を基本素性として学習データを生成している。学習データは以下の図 2 のような素性となる。

出現 書字形	出現 発音形	語彙素 読み	語彙素	原文 文字列	品詞	古語 品詞	活用形	語彙素 番号

図 2　学習データの素性

学習データを生成する過程で、歴史コーパス中の各語に関して、素性がすべて同一になる語は一度しか出現しないようにした。これは、歴史コーパスにおいて、特定のある同一単語が何度も出現することが多く、このことにより、LIBLINEAR を用いた 5 分割の交差検定の正解率に少なからず影響を及ぼす可能性が予測されるためである。

３.３　単語の分散表現を用いる手法

本研究では、3.2 節の基本素性に次の 2 つの素性を加えて学習データを生成して実験を行う。

ひとつは、各単語が現代語の語義分類において単義語であるか、多義語であるか、現代語の語義分類に存在しない単語（固有名詞など）であるかの 3 種類を示した素性である。もうひとつは、別の歴史コーパスから、word2vec（以下 w2v）を用いて 50 次元、100 次元、200 次元の分散表現のベクトルを生成し、歴史コーパス中の各単語に素性として付与したものである。

本研究において、歴史コーパスから w2v を用いてベクトルを生成する方法を説明する。まず、使用する歴史コーパスから見出し語を抜き出し、各語をすべて繋げた文章を MeCab を用いて分かち書きする。分かち書きしたデータに w2v を用いて、コーパス中宇の各語について、パラメータを変えて 50 次元、100 次元、200 次元の 3 種類の分散表現のベクトルのデータをそれぞれ得る。これらの分散表現のベクトルを、歴史コーパスから生成した学習データの各語について、ベクトル表現が存在する語のみそのベクトルを素性として追加し、ベクトル表現が存在しない語には零ベクトルを付与する。この学習データを単語の分散表現のベクトルを用いた学習データとした。単語の分散表現のベクトルを付与した学習データの素性を、以下の図 3 に示す。

出現 書字形	出現 発音形	語彙素 読み	語彙 素	原文 文字列	品 詞	古語 品詞	活用 形	語彙素 番号	**単義語か 多義語か**
								＋	**50 次元 or 100 次元 or 200 次元のベクトル**

図 3　分散表現のベクトルを加えた学習データ

以上の 2 種類の学習データに、LIBLINEAR を用いて 5 分割の交差検定を行い、正解率を調査する。

４．分散表現を利用した歴史コーパスにおける未知語義判定システムの実験

４.１　実験のデータ・設定

本実験では、歴史コーパスとして方丈記、竹取物語、虎明本狂言鬼小名、土佐日記、徒然草を使用した。この歴史コーパスの 9 つの要素を素性として、LIBLINEAR に用いるための学習データを生成した。

また、各語に対して w2v を用いた分散表現を得るための別の歴史コーパスとして、方丈記、竹取物語、虎明本狂言、土佐日記、徒然草の 5 つのコーパスをひとつに繋げて用いた。このコーパスから出現書字形の要素を抜き出し、MeCab を用いて分かち書きを行う。分かち書きしたデータに w2v を用いて、50 次元、100 次元、200 次元のベクトル表現のデータをそれぞれ得る。これらのベクトルを、歴史コーパスから生成した学習データの各語について、ベクトル表現が存在する語のみそのベクトルを素性として追加し、ベクトル表現が存在しない語には零ベクトルを付与する。以上のデータから、単語の分散表現を用いた学習データを生成した。

本実験では、word2vec による分散表現を使用した。以下の表 1 のパラメータで学習を行い、分散表現のベクトルのデータを得た。

表 1　w2v の学習パラメータ

C-BoW or skip-gram	-cbow	0
次元数	Size	50, 100, 200
ウィンドウ	-window	5
ネガティブサンプリング数	-negative	0
階層化 1:使用,0:未使用	-hs	1
最低頻度閾値	-sample	1e-3
バイナリデータ化	-binary	1
スレッドの個数	-thread	10

表 1 のパラメータにおいて、次元数の欄の 50、100、200 の 3 種類の数字は、それぞれ 50 次元、100 次元、200 次元に設定してそれぞれ 3 種類のベクトルのデータを得たことを示している。

本実験では、LIBLINEAR を用いて正解率を調査する。LIBLINEAR は、データを線形分離するための、機械学習において広く使用されているライブラリである。分類問題に対して、データに与えられた素性の情報から、線形カーネルを用いることで分類を行い、その正解率を算出する。本実験では、生成した学習データについて、それぞれ 5 分割の交差検定で学習を行い、正解率を調査した。

本手法を評価するにあたって、baseline を図 2 の素性から生成した学習データでの正解率の 77.391%とし、その学習データに、各語が単義語か多義語かの素性を加えた学習データ、さらに分散表現のベクトルを加えた学習データを用いる本手法との正解率を比較し、評価及び考察を行う。

また、w2v を用いて生成したそれぞれ 50 次元、100 次元、200 次元のベクトルを付与した学習データで実験を行い、次元数による正解率の違いについても確認し、評価および考察を行う。

４.２　実験結果

まず、baseline のデータに、各語が現代語において単義語か多義語かを示す素性を付与し

たデータでの実験結果は 79.713%となり、baseline よりもわずかに正解率が上昇した。次に、w2v で生成した 50 次元、100 次元、200 次元の分散表現のベクトルを付与したそれぞれのデータでの実験結果を以下の表 2 に示す。

表 2　ベクトルを付与した学習データの正解率（%）

baseline	単義語か 多義語か	50 次元の ベクトル付与	100 次元の ベクトル付与	200 次元の ベクトル付与
77.391	79.713	80.191	80.396	**80.533**

表 2 から、50 次元のベクトルを付与したデータでは 80.191%、100 次元のベクトルを付与したデータでは 80.396%、200 次元のベクトルを付与したデータでは 80.533%と、付与するベクトルの次元数が大きくなるほど、正解率が上昇することが確認できた。

５　考察

歴史コーパスのデータに w2v を用いて生成した分散表現のベクトルを加えた学習データを用いた実験では、分散表現のベクトルを素性として用いることにより、正解率が上がり、また、ベクトルの次元数が大きいほど、さらに正解率が高くなることが確認できた。これは、歴史コーパスの未知語義の検出に関して、分散表現を用いることの有効性を示している。

本実験では、分散表現を用いたすべてのデータで baseline を上回る正解率を得ることができたが、その上昇率はわずかであった。その要因として、分散表現のベクトルを得るために用いた歴史コーパスは、学習データに用いた歴史コーパスとは違う時代のものも入っており、時代の違いによる語義の違いによって正解率が落ちた可能性が考えられる。そこで、分類タグ付きコーパスの総データ量を増やすことや、また、同時代に編さんされた別の古文コーパスを用いて、より精度の高い分散表現を生成するなど、今後より正解率を上げるための作業および調査がさらに必要であると考える。

また、追加実験として、学習データの素性のうち、先頭三つの主要なの素性（品詞、形態素など）が正解率にどの程度貢献しているのかを実験し、調査した。学習データに用いていた 9 つの基本素性のうち、先頭の主要な素性である、出現書字形、語彙素、品詞の 3 つ素性を用いて、単義語か多義語かの素性、それぞれの次元数の分散表現のベクトルを加えて、先頭三つの素性のみでの学習データを作成した。追加実験の結果を以下の表 3 に示す。

表 3　先頭三つの素性のみの学習データの正解率（%）

	単義語か多義語か	50 次元の ベクトル付与	100 次元の ベクトル付与	200 次元の ベクトル付与
先頭三つの 素性のみ	79.131	79.843	79.843	79.487

表 3 から、先頭三つの素性のみの学習データでの正解率の方が、全ての素性を利用した場合の正解率より下がったことが確認できる。また、50 次元、100 次元の分散表現のベクトルの付与したデータよりも、200 次元の分散表現のベクトルを付与した場合のデータの方が正解率が下がっている。これは、先頭三つの素性の素性のみが必ずしも正解率に大きく寄与しているわけではないことを示している。

６　終わりに

本研究では、歴史コーパスに、単語の分散表現のベクトルを用いて、現代語の分類に当てはまらない未知語義の検出を行い、その種類を分類するシステムを提案した。実験の結果から、歴史コーパスにおける未知語の検出と分類において、分散表現のベクトルを用いること

の有効性、また、ベクトルの次元数が高いほど正解率が上昇することが示された。
また、追加実験の結果から、この未知語を検出するための学習データにおいて先頭三つの素性以外の情報も先頭三つの素性と同様に正解率に貢献していることが確認できた。

謝　辞

本研究は国立国語研究所のプロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」および「通時コーパスの構築と日本語史研究の新展開」への関連研究の一部として研究成果を報告したものである。

文　献

Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki and Hiroyuki Shinnou(2018), “All-words Word Sense Disambiguation Using Concept Embeddings”, LREC 2018, no 100,..

遊佐宣彦・佐々木稔・古宮嘉那子・新納浩幸(2017)「分散表現に基づく日本語語義曖昧性解消における類義語と辞書定義文を併用した語義表現の有効性」　言語処理学会第 23 回年次大会発表論文集, pp. 82-85.

新納浩幸・佐々木稔(2012)「外れ値検出手法を利用した新語義の検出」　自然言語処理 19 巻 5 号, pp. 304-327.

宮島達夫・石井久雄・安部清哉他(2014)『日本古典対照分類語彙表』　笠間書院.

小木曽智信(2011)「通時コーパスの構築に向けた 古文用形態素解析辞書の開発」　情報処理学会研究報告, pp. 1-4.

Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig(2013). “Linguistic Regularities in Continuous Space Word Representations”, In Proceedings of NAACL 2013, pages 746–751..

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean(2013). “Efficient Estimation of Word Representations in Vector Space”, In Proceedings of ICLR Workshop 2013, pages 1–12..

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean(2013). “Distributed Representations of Words and Phrases and their Compositionality”, In Proceedings of NIPS 2013, pages 1–9..

国立国語研究所（2018）『日本語歴史コーパス』（バージョン 2018.3，中納言バージョン 2.4.2）https://chunagon.ninjal.ac.jp/ (2018 年 4 月 30 日確認)

国立国語研究所 (2017)　分類語彙表－増補改訂版データベース
http://pj.ninjal.ac.jp/corpus_center/goihyo.html (2018 年 1 月 9 日確認)

LIBLINEAR -- A Library for Large Linear Classification
https://www.csie.ntu.edu.tw/~cjlin/liblinear/ (2017 年 12 月 16 日確認)

MeCab: Yet Another Part-of-Speech and Morphological Analyzer(2017)
http://taku910.github.io/mecab/ (2017 年 12 月 16 日確認)

国立国語研究所(編) (2017) 『現代日本語書き言葉均衡コーパス(BCCWJ)』
http://pj.ninjal.ac.jp/corpus_center/bccwj/ (2017 年 12 月 26 日確認)