

国立国語研究所学術情報リポジトリ

単義語と共起する多義語に対する分散表現を利用した語義分析

メタデータ	言語: Japanese 出版者: 公開日: 2018-03-20 キーワード (Ja): キーワード (En): Balanced Corpus of Contemporary Written Japanese (BCCWJ) 作成者: 遊佐, 宣彦, 佐々木, 稔, 古宮, 嘉那子, 新納, 浩幸, YUSA, Yoshihiko, SASAKI, Minoru, KOMIYA, Kanako, SHINNOU, Hiroyuki メールアドレス: 所属:
URL	https://doi.org/10.15084/00001522

単義語と共起する多義語に対する 分散表現を利用した語義分析

遊佐 宣彦 (茨城大学大学院理工学研究科情報工学専攻)¹

佐々木 稔 (茨城大学工学部情報工学科)

古宮 嘉那子 (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)

Sense Analysis for Polysemous Words That Co-occur with Monosemous Words Using Word Embeddings

Yoshihiko Yusa (Ibaraki University)

Minoru Sasaki (Ibaraki University)

Kanako Komiya (Ibaraki University)

Hiroyuki Shinnou (Ibaraki University)

要旨

多義語の語義曖昧性解消 (Word Sense Disambiguation, WSD) を行う際、一般的な手法では周辺の共起単語を特徴として利用する。しかし、周辺の単語には多義語が多く含まれるため、正確な語義識別が困難である。よって本研究では、単義語と共起関係にある多義語の語義分布について分析を行い、単義語と共起する多義語がそれぞれ特定の語義を持ちやすいのかを \cos 類似度と k -means クラスタリングアルゴリズムを用いて調査した。初めに、単義語に共起する多義語のベクトル分布を調査した結果、多義語ベクトルは同様の意味を持つ単語同士で固まる傾向があり、単義語と共起する多義語のベクトルは語義識別に有効であることが分かった。ただし、多義語の種類によっては語義を確定する情報として不足する可能性があることも分かった。次に、多義語に共起する単義語のベクトル分布を調査した。その結果、単義語ベクトルは同様の意味を持つ単語同士で固まる傾向があることがわかり、一部の単義語クラスは語義識別に有効であることがわかった。しかし、共起頻度が高い単義語でもその種類によっては多義語の語義を確定するには至らないものがあることも分かった。ただし、このような単義語については文脈内の他の単義語などの情報を加えることによって補える可能性がわかったため、この点はさらに検証を続けていく必要がある。今後は、単義語-多義語ベクトルに文脈内の他の単語の情報を追加した場合の語義識別の検証を行うとともに、単義語と多義語の共起情報を使用した実際の WSD 手法を実現したいと考えている。

1.はじめに

多義語の WSD を行う際、一般的な手法では周辺の共起単語が特徴として利用される。しかし、周辺の単語には多義語が多く含まれるため、正しい語義識別ができない問題がある。また、近年 WSD の分野では word2vec を用いた単語の分散表現の研究が数多く行われている。word2vec からは大量の文章データを基に単語間の意味関係をベクトルとして得ること

¹ 16nm726a@vc.ibaraki.ac.jp

ができ、WSD においても有用なデータを得られることが期待されている(佐々木ほか(2016))。しかし、word2vec では文章中に出現する単語を対象にベクトルを生成するため、多義語のベクトルは複数の語義の特徴を含んでしまい、語義の特徴を分割して抽出できない問題がある。

そこで本研究では、単義語の単語ベクトルに注目した。単義語は語義を一つしか持たないため、その単語ベクトルは複数の語義の特徴を含まず一意に決定される。このことから単義語の単語ベクトルは多義語の語義識別に有効だと考えられる。この仮定を検証するために、本研究では単義語と共起関係にある多義語の語義分布について分析を行う。現代日本語書き言葉均衡コーパスを用いて、初めに出現頻度の高い単義語に共起する多義語について分析する。次に Semeval2010 日本語タスクで課題とされた多義語を対象に共起する単義語の分布を分析する。この分析では、単義語と共起する多義語がそれぞれ特定の語義を持ちやすいのかについて調査する。

さらに、分散表現を利用して「単義語から見た多義語」と「多義語から見た単義語」のベクトルがどのような特徴を持つのか分析する。単義語は複数のベクトルに分かれず、ひとつのベクトルで表現されるものと仮定し、単義語と頻繁に共起する多義語がどのような語義の特徴を持つのか調査する。

既存手法(Li et. al. (1999))では中国語の WSD における単義語と多義語の共起情報の有効性を検証しているが、日本語の場合や分散表現を利用した場合の有効性は検証されていない。既存手法では登録された単義語が文中にある場合のみでしかその特徴を利用できないが、分散表現を利用できるようにした場合、単義語の情報を他の単語の特徴と比較することで登録している単義語が含まれない文章の語義識別も行えるようになる。したがって日本語における単義語と多義語の共起ベクトルの語義識別への有効性の分析は有益だと考える。

2.分析方法

本研究では、単義語からみた多義語の分布と多義語からみた単義語の分布の二つについて分析する。

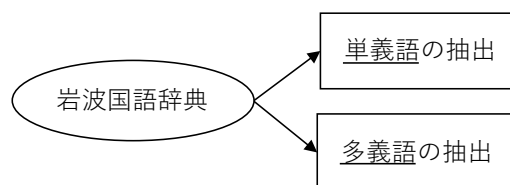


図1：単義語と多義語の抽出

二つの分析を行う前に、辞書から単義語の抽出を行う。今回は、岩波国語辞典から、大分類に対応する語義番号の2番以降が全て0(XXXX-0-0-0)かつ語義がただ一つのものであるだけを選び、それを単義語とした。一方、多義語については辞書内に含まれる単語のうち「今回単義語と定義した単語以外の単語」として定義した。

また、本研究で扱う単義語と多義語はすべて形態素解析器 MeCab で名詞と判断される名詞のみに絞って実験を行う。

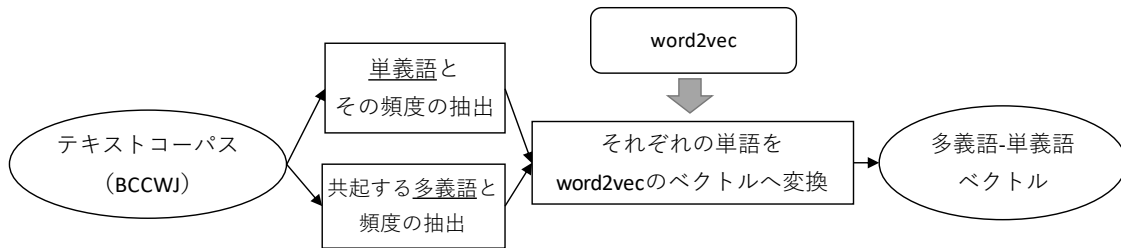


図2：単義語と共起する多義語の分析手法

2.1 単義語と共起する多義語の分析

まず、単義語と共起する多義語の分析を行うことで、単義語と共起する多義語の分布の特徴を調査する（図2）。

事前に単義語と多義語を辞書から取得した後、テキストコーパス内での単義語の出現頻度を計算する。そのうち最も頻度が高いものから100単語を対象とし、それぞれの単義語について共起する単語の抽出を行った。共起単語として抽出する範囲は単義語の前後2単語とし、その中から出現する多義語の頻度を計算した。

共起する多義語の頻度を計算後、共起頻度が50以上の単語組に絞って分析する。本研究では word2vec を用いて作成したベクトルデータとして、国立国語研究所から配布された nwjc2vec(浅原ほか (2017))を使用し単義語と多義語のベクトルを取得した。取得した多義語ベクトルから単義語のベクトルを減算することで多義語-単義語のベクトルを作成する。

作成したベクトルは、cos 類似度を用いて、どのような分布になっているかを分析した。

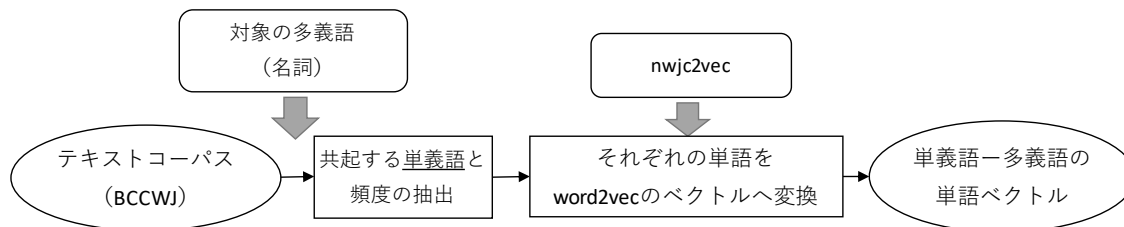


図3：多義語と共起する単義語分布の分析手法

2.2 多義語と共起する単義語分布の分析

次に、多義語と共起する単義語のベクトル分布の分析を行う。ここでは、多義語からみた単義語の分布がどのようなになっているかを調査する(図3)。

まず、Semeval2010 で公開された課題データから品詞が名詞である単語を対象単語として抽出する。次に、テキストコーパスから対象の多義語の前後2単語を共起単語とし、その中に存在する単義語の頻度を計算した。

共起する単義語の頻度を計算後、共起頻度について上位50の単語組に絞って分析を行う。2.1と同様に nwjc2vec のデータを使用し、単義語のベクトルから多義語のベクトルを減算することで単義語-多義語のベクトルを作成する。

作成したベクトルは、k-means クラスタリングアルゴリズムを用いて、どのような分布になっているかを分析した。

3.使用データ

本実験で使用するデータは、国立国語研究所による現代日本語書き言葉均衡コーパス (BCCWJ Version 1.1) から Disc2 の文章データと Semeval2010 日本語 WSD タスクで課題として公開されたデータを利用する。Semeval2010 では 50 個の対象単語とその単語を使用した用例文の文章データ 50 個が用意されており、今回はその中から名詞の単語 22 個に絞って実験を行った。

また、分散表現のデータとして、国語研究所が word2vec を用いて作成した nwcj2vec を使用した。さらに k -means クラスタリングを行う際に python のライブラリである scikit-learn を用いた。

4.実験

単義語と共起する多義語の分布を分析するために、単義語と共起する多義語ならびに多義語と共起する単義語のそれぞれを分散表現で表し、その分布を cos 類似度と k -means クラスタリングで評価する。

表 1: 「企業」(単義語) と共起する多義語のベクトル類似度表

多義語	cos類似度の高い多義語 (上位3つ)		
小	大	中	ファミリー
中	内	間	共
大	小	中	団
経営	事業	関係	代表
民間	行政	機関	地方
者	人	等	士
等	他	場合	一部
庁	機関	団	行政
内	中	内部	等
間	中	共	先

4.1 単義語と共起する多義語の分布

多義語-単義語のベクトルの類似度を同単義語内で比較した例を表 1 に示す。「多義語」は単義語との共起頻度が高い多義語から降順に記されており、「cos 類似度の高い多義語」では他の多義語-単義語のベクトルから cos 類似度の高い多義語 3 つを示している。

この表から、単義語からみた多義語のベクトル分布は似たような意味を持つものが固まる傾向があることがわかる。例えば「小-企業」「大-企業」や「民間-企業」「行政-企業」、「等-企業」「他-企業」などは高い類似度を示した。特に、「小-企業」「大-企業」の組について見ると「企業」と共起した際の「小」の語義と「大」の語義は二つとも【規模・程度】を表す意味で定まる傾向にあることがわかった。

しかし、「中-企業」について見ると、このベクトルは「小-企業」「大-企業」という【規模・程度】を表す意味の固まりから高い類似を持たれている一方で、「内-企業」「間-企業」のような【内部・含有状態】を示す意味の単語と高い類似を持っている (図 4)。つまりこのことは、「企業」と共起している「中」の語義を識別する際に、「小」「大」の意味の塊と判断するか、もしくは「内」「間」としての意味の塊と判断するか二通りの判断ができることを表しており、語義は一意に決まらないことがわかる。

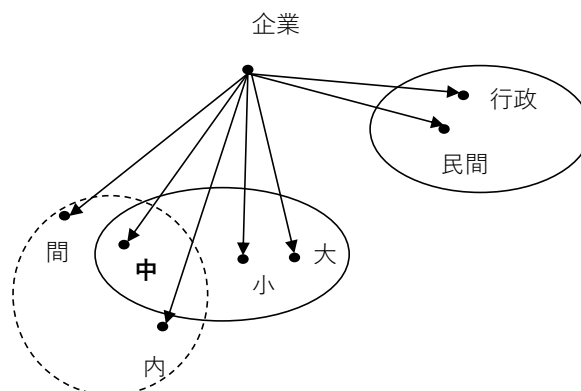


図4：多義語-単義語ベクトル分布のイメージ「企業」

これらのことから、単義語と多義語の共起の情報は多義語の意味の判別に有効である一方で、多義語の種類によっては単義語との共起にかかわらず意味ドリフトが起こってしまい語義を確定できない可能性があることが分かった。

4.2 単義語と共起する多義語の分布

単義語-多義語のベクトルを k -means でクラスタリングした結果($k=10$)を表2に示す。「クラスタ id」は k -means アルゴリズムで割り振られたクラスタ番号を表し、「単義語一覧」ではクラスタ内に属する単義語 (全 50 個) の分布を示している。

表2: 「意味」(多義語) と共起する単義語ベクトルのクラスタリング結果 ($k=10$)

クラスタid	0	1	2	3	4	5	6	7	8	9
単義語一覧	本来	存在	単語	質問	今一	両方	重要	特別	全て	変化
	厳密	解釈	記号	確認	何処	レベル	重大	普通	自由	転換
		区別	漢字	説明	何方	機能	価値	通常	完全	作用
		象徴	用語	発見		部分	歴史	同様		
		政治	言語			構造	新た	使用		
		否定	諺				明確			
		行動					必要			
		戦争								
		表現								
		本質								
	主観									

この表から、多義語と共起する単義語のベクトル分布は、類似の意味を持つ単義語で固まる傾向にあることがわかる。例えばクラスタ id が 2 の単義語をみると、「単語」「記号」「漢字」「用語」「言語」「諺」といった【言葉】に関する意味を持つものでクラスタを作っている。「意味」の語義は表3のように三つ存在するが、【言葉】に関する語義は一つであり、これらの単義語と共起時はこの語義に定まる傾向があることが分かった。

表 3:岩波国語辞典における「意味」の語義番号と定義文

Headword いみ 【意味】
2843-0-0-0-0 ((名・ス他))
2843-0-0-1-0 <1>その言葉の表す内容。意義。「辞書を引けば一がわかる」
2843-0-0-2-0 <2>表現や行為の意図・動機。「どうい—でそんなことをしたのか」
2843-0-0-3-0 <3>表現や行為のもつ価値。意義。「そんな事をしても一がない」

一方で、クラスタ id が 7 の「特別」という単語について「意味」との共起した場合の用例を見ると、「(単語の) 特別な意味」「(この行動に) 特別な意味 (はない)」「(国にとって) 特別な意味 (を持つ)」のように「特別」と共起していても「意味」の語義は確定しないことがわかった。また、クラスタの取り方によって「使用」「特別」などのほぼ関係のない意味の単語が同クラスタ内に分類されることもあることがわかる。

ただし、上記の「特別」の用例では、「特別」「単語」「意味」や「行動」「特別」「意味」などの組み合わせで見ると意味が定まると考えられる。よって、単義語-多義語の一組の情報でのみ語義を特定するのではなく、この情報を手掛かりとして文脈内にある他の単語と組み合わせることでより識別精度が高くなりそうなことが推測できる。また、クラスタの分け方については今回の実験とは別に検証を進めていく必要がある。

5.おわりに

本研究では、単義語と多義語の共起関係とそのベクトルについて分析し、語義識別に有効であるかどうかを調査した。初めに、単義語に共起する多義語のベクトル分布を調査した結果、多義語ベクトルは同様の意味を持つ単語同士で固まる傾向があり、単義語と共起する多義語のベクトルは語義識別に有効であることが分かった。ただし、多義語の種類によっては語義を確定する情報として不足する可能性があることも分かった。次に、多義語に共起する単義語のベクトル分布を調査した。その結果、単義語ベクトルは同様の意味を持つ単語同士で固まる傾向があることがわかり、一部の単義語クラスタは語義識別に有効であることがわかった。しかし、共起頻度が高い単義語でもその種類によっては多義語の語義を確定するには至らないものがあることも分かった。ただし、このような単義語については文脈内の他の単義語などの情報を加えることによって補える可能性がわかったため、この点はさらに検証を続けていく必要がある。

今後は、単義語-多義語ベクトルに文脈内の他の単語の情報を追加した場合の語義識別の検証をしていくとともに、単義語と多義語の共起情報を使用した実際の WSD 手法を実現したいと考えている。

謝辞

本研究の一部は国立国語研究所の共同研究プロジェクト「all-words WSD システムの構築及び分類語彙表と岩波国語辞典の対応表作成への利用」の研究成果を報告したものである。

参考文献

- 佐々木稔・古宮嘉那子・新納浩幸(2016). 「分散表現に基づく日本語語義曖昧性解消における辞書定義文の有効性」言語処理学会第22回年次大会発表論文集, P11-1, pp.449-452.
- Li Juanzi, and Huang Changning(1999). “A Model for Word Sense Disambiguation”
Cmputational Linguistics and Chinese Language Processing Vol.4, No.2, pp.1-20.
- 浅原正幸・岡照晃(2017). 「nwjc2vec『国語研日本語ウェブコーパス』に基づく単語の分散表現データ」言語処理学会第23回年次大会講演論文集, E1-5, pp.82-85.