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

佐々木 稔　　古宮 嘉那子　　新納 浩幸
茨城大学工学部情報工学科
{minoru.sasaki.01, kanako.komiya.nlp, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

1 はじめに

近年，語義曖昧性解消 (WSD) において単語の分散表現を利用した研究が数多く行われている．単語の分散表現は mikolov らの手法 [3] に代表されるようなディープラーニング技術を用いて単語間の意味的な関係をベクトルで表現したものである．この分散表現を使うことで，意味分類した既存のシソーラスを使わずに，大量の文書データから単語の概念関係が得られるために，WSD 以外の自然言語処理タスクでも利用されている．

分散表現を利用した WSD には，従来の語義を識別する対象単語を指定した WSD と指定しない All-words WSD の 2 種類が存在するが，本稿では対象単語を指定した WSD を対象とする．この場合も一般的な WSD 手法と同様に，教師あり手法と教師なし手法の 2 種類が存在する．教師あり手法は用例文中の単語に対して分散表現を求めて概念をベクトルとして表す手法 [4] や特徴空間を拡張する手法 [6][7] が存在する．教師なし手法は辞書の定義文から語義の分散表現を求めて，テストデータとの比較によって語義識別を行う [1]．

本稿では，日本語辞書記述の利用に着目し，辞書の定義文から得られた語義の分散表現が WSD に有効であるかどうか検証する．教師なしの既存手法は英語辞書を用いているため，日本語辞書を利用した場合の有効性は検証されていない．また，教師ありの既存手法では辞書の情報が使われていないため，辞書情報を追加した場合の WSD への有効性も検証されていない．したがって，日本語辞書情報を加えた分散表現に基づく WSD の有効性検証は有益だと考えられる．教師なし，教師ありのそれぞれについて WSD システムを構築し，日本語辞書情報の有効性を検証する．

2 単語分散表現の作成

2.1 使用データ

単語の分散表現を求めるためのデータとして，日本語 Wikipedia の文書を使用する [1]．文書データに対し，形態素解析器 MeCab[2] を用いて単語列に分割する．このとき，動詞などの活用はすべて見出し語に変換する．使用する PC の環境に制限により，得られた単語列をすべて利用して分散表現を求めることができなかったため，単語列ファイルの先頭から約 2.14GB 分のデータを使用することとした．

2.2 単語分散表現の作成

単語列データから単語の分散表現を求めるために，word2vec[3] というツールを利用した．word2vec を実行する際，学習モデルとして Continuous Bag-of-Words(C-BoW)，分散表現の次元数を 200，ウインドウ幅を 5，ネガティブサンプルを 5 として学習を行った．その結果，379,726 単語の分散表現が得られた．

3 教師なし WSD

日本語辞書情報を用いた分散表現に基づく WSD 手法について説明する．この手法と有効性を比較するために，一般的な教師なし手法である Lesk アルゴリズムに基づく WSD 手法についても説明する．

3.1 単語分散表現を用いた WSD

単語分散表現を用いた WSD は，語義を分散表現で表し，入力文の分散表現と比較することで語義を識別

[1] 2015 年 10 月 2 日時点でのダンプデータ
[2] http://mecab.googlecode.com/svn/trunk/mecab/doc/
[3] https://code.google.com/p/word2vec/

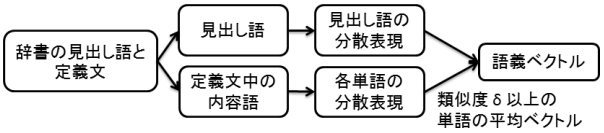

図 1: 語義の分散表現の求め方

する方法である．語義の分散表現は図 1 のように辞書の定義文から求める．まず，各語義の定義文に対して，mecab を利用して対象単語以外の名詞，動詞，形容詞，副詞の内容語を抽出する．次に，対象単語の見出し語と辞書定義文の内容語をそれぞれ word2vec で求めた分散表現に変換して，ベクトルで表現する．見出し語と各内容語のコサイン類似度を計算し，閾値 δ 以上を持つ内容語を語義の分散表現を作るための候補集合に加える．候補集合への追加判定を行った後，候補集合にある内容語の分散表現を平均したベクトルを語義の分散表現として出力する．この処理を対象単語の各語義に対して行う．

対象単語の語義を識別する際は，まず入力となる用例文から文脈ベクトルを作成する．用例文から mecab により内容語を抽出し，内容語の分散表現の平均を文脈ベクトルとする．文脈ベクトルと各語義の分散表現に対して，コサイン類似度を計算し，類似度が最大となる語義を識別結果として出力する．

3.2 階層型 Lesk アルゴリズムを用いた WSD

教師なし WSD の古典的な手法として，Lesk アルゴリズムがある [2]．このアルゴリズムは入力された用例文と定義文の間に，重複する単語が最も多い語義を識別結果として出力する．しかし，Lesk アルゴリズムは対象単語の定義文に一致する単語が用例文に存在しないことが頻繁に起こるために識別に失敗する事がある．この問題を解消する手法はこれまでに研究されているが，本稿では階層型 Lesk アルゴリズムを構築し，用例文と辞書定義文の概念的な重複を捉えることとする．

階層型 Lesk アルゴリズムでは，入力された用例文と定義文の単語対に対して，Lesk アルゴリズムでスコアを計算する．各単語対の定義文で重複する単語数をスコアとして，それをすべての単語対に対して計算し，用例文と語義の類似度とする．最終的に最も高い類似度を持つ語義が識別結果として出力される．ただし，ここで扱う単語は 2 文字以上の単語に限定して処理を行った．

この方法を利用する場合，重要な単語が一致するだけではなく,「する」や「こと」といった，どの定義文でもよく使われる単語の一致が頻繁に起こる．そのため，単語対の数が多いほど類似度が高くなる傾向があり，記述が長い語義ほど識別結果として選ばれやすくなる．この問題を解消するため，どの定義文でも使われる一般的な語である { する，いる，こと，もの，せる，よう，れる，なる，そこ，それ，られる } を定義文から除いて Lesk スコアを計算した．

4 教師あり WSD

教師あり WSD においても，日本語辞書情報の有効性を検証するため，辞書定義文に訓練データを追加する方法を説明する．この方法の適合率を比較するため，定義文を使わずに訓練データのみを利用した方法についても説明をする．

4.1 定義文に訓練データを追加する方法

辞書の定義文には語釈文の他にいくつかの用例文が記載されている場合がある．この用例が多いほど語義の特徴が明らかになり，語義を理解しやすくなると考えられる．また，語義によっては定義文が短い場合があり，語義の特徴を捉えにくいという問題も存在する．これらの問題に対して，語義が明らかとなっている訓練データの用例文を語義の定義文に追加し，3.1 節と同じ方法で語義の分散表現を作成することを考える．得られた各語義の分散表現と入力文の文脈ベクトルのコサイン類似度を計算し，類似度が最大となる語義を識別結果として出力する．

4.2 各用例文の分散表現を利用する方法

辞書情報の有効性を評価するために，辞書情報を利用せずに訓練データだけを利用する方法を考える．訓練データの各用例文に対して分散表現を求め，ラベル付き文脈ベクトルを作成する．対象単語の語義を識別する際は，入力の用例文から文脈ベクトルを求め，訓練データから得られたラベル付き文脈ベクトルとのコサイン類似度を計算し，最も高い値を持つベクトルのラベルを識別結果として出力する．

5　実験

日本語辞書情報の有効性を検証するために，上記の教師なし，教師あり WSD 手法を用いた実験を行う．

5.1　データ

本実験で使用するデータは，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．これは 50 語の対象単語が指定され，その各対象単語について訓練データとテストデータが存在する．訓練データには対象単語の語義ラベルが付与された 50 文の用例文が存在する．テストデータにも 50 文の用例文が存在し，システムが各文の対象単語の語義を識別し，語義の適合率を評価する．

6　実験結果

Semeval2010 データについて，教師なし，教師あり WSD 手法を利用して語義識別実験を行った．この実験では，階層型 Lesk アルゴリズム以外の 3 つの方法は分散表現を求める際にパラメータ δ が必要となる．今回の実験では，教師なし WSD の平均適合率が最も高い結果を示した $\delta = 0.0$ にパラメータの値を固定して処理を行った．

各 WSD 手法を用いた実験結果を表 1 に示す．表 1 における「Lesk」は階層型 Lesk アルゴリズムを用いた方法,「教師なし」は単語分散表現を用いた方法,「定義＋用例文」は定義文に訓練データを追加した方法,「用例文」は各用例文の分散表現を利用した方法を表す．表の行は対象単語について各方法を用いた場合の適合率を表し，最後の行に平均適合率を示す．さらに，識別性能比較のため，Project Next NLP[4] において誤り分析で使われた WSD システムの適合率も示す [8]．

教師なし WSD では，単語分散表現を用いた方法の平均適合率は 52.84%で，階層型 Lesk アルゴリズムを用いた方法の 51.56%を上回る結果となった．教師あり WSD においては，定義文に用例文を追加する方法が 63.20%，各用例文を分散表現に変換した方法が 70.20%という平均適合率であった．これら 4 つの方法はシソーラスとサポートベクターマシン (SVM) を用いた Project Next NLP で使われたシステムを下回る結果となった．今回の実験では識別時にコサイン類似度を用いたが，SVM を利用すると適合率が上がる可能性がある．その実験については今後の課題とする．

[4]https://sites.google.com/site/projectnextnlp/

7　考察

日本語辞書情報のみを使った教師なし WSD は約 50%の平均適合率となり，用例文を使った WSD と比較して低い適合率であった．辞書情報はこれまでの研究 [5] でも言われるように，WSD においてある程度の効果は認められる．しかし，辞書の定義文から有効な情報の取得方法と分散表現の作成方法について，さらに改善を行う必要があると考えられる．

階層型 Lesk アルゴリズムは，適合率の高い単語と低い単語に極端に分かれる傾向がある．定義文に数多くの内容語が存在するほど高いスコアを持ち，どんな入力文も同じ語義になりやすい．対象単語の「教える」は,「さとす，戒める」という語義の定義文に他の語義よりも内容語が多く含まれ，すべてのテストデータでこの語義が選ばれたために低い適合率となっている．また,「求める」では各語義の定義文の長さはほとんど同じであるが,「手に入れる」の語義にある「手」の定義文に内容語が多く含まれていたために，識別誤りが多い結果となった．逆に,「経済」や「文化」といった単語はテストデータにおいて多数を占める語義が選ばれやすいため，高い適合率となっている．

単語分散表現を用いた方法は，辞書の定義文に内容語が多く含まれ，その語義がテストデータでよく使われていれば，適合率が高くなる傾向がある．例えば，対象単語「会う」は「対面する，会見する」の語義を定義文の記述が多い「釣り合う，調和する」の語義に誤って識別し，適合率が低い結果となった．対象単語「取る」は Project Next NLP システムでは「負う，引き受ける」の語義がすべて誤りであったが，日本語辞書情報を使うことで正しく識別できた．日本語辞書情報を使うことで，出現頻度の少ない語義について正解語義を識別しやすくなる効果があることが分かった．また，単語分散表現を用いた方法は Lesk アルゴリズムよりも高い適合率が得られたため，定義文中の単語の重複を使うよりも，単語の分散表現を使って単語の関連性を測る方が効果が大きいと考えられる．

辞書定義文に訓練データを追加する方法は，日本語辞書情報だけを使う場合と比較して約 10%高い平均適合率が得られた．分散表現を利用しているため，類似した単語は類似したベクトルに変換されるが，辞書定義文に数多くの用例文が含まれる方が語義の特徴を捉えやすい事が分かる．しかし，用例文を含めることで識別誤りが増える対象単語も存在した．例えば,「求める」では「買う，手に入れる」の語義と「要求する」の語義に共起する単語が類似したため，識別誤りが増加する結果となった．複数の語義でほとんど同じ共起

表 1: 各手法による WSD 実験の結果

単語	Project Next NLP	Lesk	教師なし	定義＋用例文	用例文
相手	80	14	46	82	78
会う	92	18	18	38	48
あげる	52	36	28	30	42
与える	70	58	30	58	56
生きる	94	10	22	72	88
意味	44	24	46	46	48
入れる	74	72	48	54	66
大きい	94	94	52	74	86
教える	52	0	80	70	60
可能	64	56	56	56	54
考える	98	92	70	90	98
関係	96	78	60	70	80
技術	82	84	84	74	76
経済	98	98	82	88	90
現場	76	22	24	72	76
子供	62	64	62	62	56
時間	84	66	40	82	82
市場	56	70	70	54	68
社会	86	54	74	66	76
情報	84	84	70	82	78
勧める	92	48	52	72	66
する	72	0	16	44	38
高い	88	84	74	80	86
出す	50	56	44	40	46
立つ	50	8	38	38	50
強い	90	92	80	66	80
手	78	74	74	68	76
出る	52	58	32	60	56
電話	78	46	60	84	72
取る	28	22	40	32	32
乗る	78	10	38	60	70
場合	84	58	50	60	72
入る	56	50	58	66	56
はじめ	88	36	44	70	76
始める	86	42	60	56	68
場所	96	4	16	66	96
早い	70	48	72	72	74
一	90	86	46	64	88
開く	84	90	40	56	92
文化	98	98	70	94	98
他	100	86	70	64	98
前	76	38	34	48	68
見える	70	54	28	50	46
認める	82	14	46	62	68
見る	78	78	62	58	70
持つ	80	64	42	58	68
求める	76	4	70	38	62
もの	88	12	66	72	84
やる	96	54	84	80	94
良い	54	70	74	62	54
平均適合率	76.92	51.56	52.84	63.20	70.20

単語を持つと語義の識別は非常に難しくなるため，定義文のみを使う方が高い適合率になったと考えられる．

各用例文の分散表現を利用する方法が 4 つの方法の中で最も高い平均適合率となった．それでも Project Next NLP システムと比較すると低い平均適合率で，共起単語の範囲や品詞情報，係り受けや識別モデルの種類などの違いにより，識別結果に違いが生じたと考えられる．そのため，識別モデルとしてサポートベクターマシンを使った実験や品詞情報を加えた分散表現の獲得手法を作成するなどの改良を行う必要がある．

8 おわりに

本稿では，分散表現に基づく日本語 WSD において，辞書の定義文から得られた語義の分散表現の有効性について検証を行った．教師なし，教師ありの WSD 手法を利用して WSD 実験を行った結果，教師なし WSD では，単語分散表現を用いた方法の平均適合率が Lesk アルゴリズムを用いた方法を上回る結果となった．日本語辞書情報を使うことで低頻度語義については正しく識別された用例が存在したことから，語義の識別にある程度の効果があることが分かった．教師あり WSD においては，定義文に用例文を追加する方法と各用例文を分散表現に変換した方法の両方で分散表現を使わない手法を下回る結果であったが，分散表現を利用する場合においても対象単語周辺の共起単語の使用が有効であることが分かった．

今後は，用例文の分散表現に辞書情報を効果的に組み込む工夫をするとともに，WSD に有効な分散表現・品詞情報を加えた分散表現の作成方法についての検討，識別モデルとしてサポートベクターマシンを使った実験などを行い，高い適合率を持つ WSD 手法を目指したいと考えている．

参考文献

[1] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1025–1035, 2014.

[2] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, pp. 24–26. ACM, 1986.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[4] 新納浩幸, 古宮嘉那子, 佐々木稔. all-words wsd のための概念辞書の自動作成. 情報処理学会自然言語処理研究会, pp. NL–224–13, 2015.

[5] 白井清昭, 八木恒和. コーパスと辞書定義文中の上位概念を用いた頑健な語義曖昧性解消. 言語処理学会第 10 回年次大会, pp. 745–748, 2004.

[6] 菅原拓夢, 笹野遼平, 高村大也, 奥村学. 単語の分散表現を用いた語義曖昧性解消. 言語処理学会第 21 回年次大会, pp. 648–651, 2015.

[7] 山木翔馬, 新納浩幸, 古宮嘉那子, 佐々木稔. 分散表現を用いた教師あり機械学習による語義曖昧性解消. 情報処理学会自然言語処理研究会, pp. NL–224–17, 2015.

[8] 新納浩幸, 村田真樹, 白井清昭, 福本文代, 藤田早苗, 佐々木稔, 古宮嘉那子, 乾孝司. クラスタリングを利用した語義曖昧性解消の誤り原因のタイプ分け. 自然言語処理, Vol. 22(5), pp. 319–362, 2015.