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

遊佐 宣彦[1]　佐々木 稔[2]　古宮 嘉那子[2]　新納 浩幸[2]
[1]茨城大学大学院理工学研究科情報工学専攻
[2]茨城大学工学部情報工学科
{16nm726a, 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(Word Sense Disabigation,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。先には、日本語辞書の語義定義文の分散表現を利用した教師ありなし学習の有効性が検証されており[1]、語義の識別への効果などが認められた。

本稿では、日本語辞書と分類語彙表を併用して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。既存手法は日本語辞書の定義文から得られた分散表現によって有効性を検証しているが、類義語を加えた場合の有効性は検証されていない。したがって、日本語辞書と分類語彙表を併用した分類表現によるＷＳＤへの有効性検証は有益だと考える。日本語辞書の定義文から得た分類表現と、日本語辞書と分類語彙表との併用で得られた分類表現のＷＳＤ実験を比較することによってそ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から「新聞」「雑誌」「書籍」の文章データを使用する。文章データは、形態素解析器 MeCab を用いて単語列に分割し、動詞などの活用語はすべて見出し語に変換する。これによって 158MB のデータを得ることができ、分散表現の作成に使用することとした。

2.2 単語分散表現の作成

上記の単語列データから単語の分散表現を求めるため、word2vec 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(CBoW)、分散表現の次元数を 200、ウインドウ幅を 6、ネガティブサンプルを 5 として学習を行い、83,025 単語の分散表現が得られた。

3 WSD 手法

日本語辞書を用いた分散表現に基づくＷＳＤと、分類語彙表を併用した分散表現に基づくＷＳＤに用いた手法について説明する。

3.1 日本語辞書に基づく WSD

このＷＳＤでは日本語辞書の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する。語義の分類表現は

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.

図 1 のように求める。まず、各語義の定義文に対して MeCab を利用して名詞、動詞、形容詞の内容語を見出し語と変換し抽出する。次に、抽出した単語列を word2vec で求めた分散表現に変換し、語義定義文の分散表現とする。

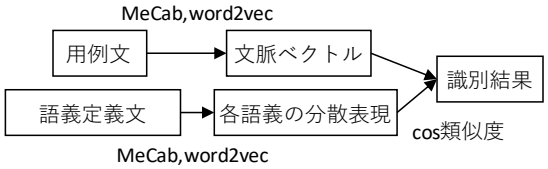

図 1:分類語彙表併用手法の手順

対象単語の語義識別の際には、入力となる用例文から文脈ベクトルを作成して比較を行う。用例文からは MeCab を利用して内容語のみを抽出し、そ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をコサイン類似度で比較し、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 分類語彙表を併用した WSD

日本語辞書の語義を基に、分類語彙表からそれぞれの語義に対して類義語を手動で紐づけした。また、類義語だけでは表現しきれない語義、特に慣用表現を表す語義などに対しては、類義語の他、語義の定義文中の有効性の高そうな単語も語義として用いることで語義表現を補った。さらに、これらは 3.1 での語義表現のように一つの語義に一つの分類表現を作成するのではなく、たくさんの類義語や定義文中の単語にそれぞれ語義ラベルを紐づけて（図 2）個別で分散表現を求めた。

117-0-0-1 仲間	117-0-0-2 遊ぶ	117-0-0-3 敵手
117-0-0-1 同類	117-0-0-2 離す	117-0-0-3 好敵手
117-0-0-2 先方	117-0-0-3 争う	117-0-0-3 ライバル
117-0-0-2 相手方	117-0-0-3 敵	
117-0-0-2 向こう	117-0-0-3 対抗	

図 2: 3.2 における「相手」の語義ラベル一覧

対象単語の語義識別の際には、3.1 と同じく入力となる用例文から文脈ベクトルを作成して比較を行う。用例文の文脈ベクトルと、語義ラベルを紐づけられた各単語の分散表現をコサイン類似度で比較し、類似度が最大となる単語の語義ラベルを識別結果として出力する。（図 3）

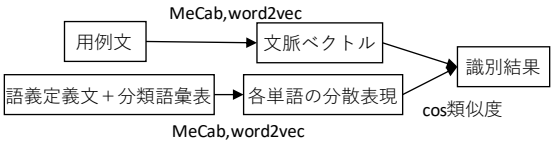

図 3:分類語彙表併用手法の手順

4 実験

日本語辞書と分類語彙表との併用で得られた分散表現におけるＷＳＤの有効性を検証するために、上記の教師なしＷＳＤ手法を用いた実験を行う。

4.1 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。データとして 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、これをテストデータに使用する。ただし、今回は類義語の有効性を検証するにあたりその中から名詞の単語 22 個に絞って実験を行った。

5 実験結果

二つの手法を用いたＷＳＤ実験結果を表

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.

表 1:二つの手法の WSD 実験の結果

語義定義文		
単語	正解数	精度(%)
相手	27	54
意味	20	40
可能	28	56
関係	22	44
技術	42	84
経済	36	72
現場	12	24
子供	31	62
時間	24	48
市場	35	70
社会	38	76
情報	33	66
手	36	72
電話	27	54
場合	33	66
はじめ	15	30
場所	12	24
一	37	74
文化	47	94
他	46	92
前	15	30
もの	37	74
平均精度		59.36

語義定義文＋類義語		
単語	正解数	精度(%)
相手	30	60
意味	23	46
可能	24	48
関係	26	52
技術	35	70
経済	36	72
現場	44	88
子供	32	64
時間	19	38
市場	30	60
社会	34	68
情報	29	58
手	29	58
電話	19	38
場合	37	74
はじめ	16	32
場所	23	46
一	8	16
文化	44	88
他	50	100
前	31	62
もの	42	84
平均精度		60.09

1 に示す。表 1 における「語義定義文」は語義定義文を分散表現化した手法、「語義定義文＋類義語」は類義語と語義定義文を併用して語義ラベルと分散表現を作成した手法を表す。表の行は対象単語についてのそれぞれ、50 個の例のうちの「正解数」と「精度」を表し、最後の行に平均精度を示す。

語義定義文と分類語彙表を併用した手法のＷＳＤでの平均精度は 60.09%で、これは語義定義文のみを用いて分散表現化する手法の平均精度 59.36%を上回る結果となった。しかし、単語ごとで見ると、語義定義文のみの精度に比べ精度が 10%以上低下している部分もあり、精度の上昇は部分的であることがわかる。

6 考察

分類語彙表の類義語と、語義定義文中の単語を利用した教師なしＷＳＤは 60.09%の平均精度となり、語義定義文をそのまま利用したＷＳＤと比較して約 0.7%の精度上昇が認められた。

今回の実験の過程で、語義定義文を利用した分散表現でのＷＳＤの傾向として、語義定義文中に含まれている単語の数が多い語義ほどテストデータの用例文との類似度が高くなり、正解として選ばれやすくなっていることが分かった。例えば「現場」では、語義 1 に含まれる単語数が 12 個に対し語義 2 の単語数は 5 個、正解として選ばれた個数は前者が 40 個に対し後者が 10 個となっている。他にも「技術」では語義 1 の単語数は 49 個に対し語義 2 の単語数は 10 個で、正解として選ばれた個数は前者が 50 個で後者が 0 個となっている。このことは精度の上昇と低下に大きく関わっており、正しく文章中の対象単語の語義を判別できない要因の一つでもある。

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.

類義語を利用した場合、すべての語義ラベルはほとんど 1 単語になっているため、単語数の偏りによる正解数の偏りは起きにくくなっている。例えば、「技術」の解答個数の内訳は、語義 1 が 35 個で語義 2 が 15 個と改善されている。「技術」ではこのことにより精度の低下を招いてしまったが、一方で「現場」や「前」では精度の向上に貢献している。

類義語を含めたＷＳＤでは、テストデータとして用意された用例文のほとんどで高い類似度を示す単語が存在することが分かった。例えば、「一」の類義語として追加した「単一」はほとんどの用例文で 0.3~0.6 の高い類似を示し、その結果「単一」にラベル付けされた語義の解答が多くなり精度が低下する結果となった。また同じく「一」では「いま一つ〜」という慣用表現としての「一」の語義定義文「物足りない」が 0.2~0.4 の高い類似を保っており精度低下の要因となっていた。このことを基に、「単一」のようなほとんどの用例文との高い類似を示す単語をラベルから削除したところ「一」の精度は 58%まで上昇し、全体の平均精度は 62%まで向上した。一方で、「いま一つ〜」のような慣用句的な使い方をする場合の語義は適当な類義語を探すことが難しく、辞書に載っている用例も少ないため、ラベルデータから単語を削除するとラベルを表す単語がほとんど無くなってしまう問題がある。このことから、慣用表現を示す語義の有用な情報抽出の方法を模索し改善を行う必要があると考えられる。

7 おわりに

本稿では、分散表現に基づく日本語ＷＳＤにおいて、辞書の定義文と分類語彙表の類義語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。辞書の定義文のみを使ったＷＳＤ手法と、分類語彙表の類義語併用のＷＳＤ手法を利用してＷＳＤ実験を行った結果、分類語彙表併用の方法の平均精度が辞書定義文のみの場合を上回る結果となった。ラベルに単語列を紐づけたものではなく、類義語や定義文中の単語一つ一つを並列化してラベルづけすることによって、単語数の偏りによる解答の偏りの改善が見られたことから、語義の識別にある程度の効果があることが分かった。また、ラベル付けされた単語の中には、ほとんどの用例文での類似度が高くなる汎用的な単語が存在し、それらを削除することによって精度が向上することが分かった。

今後は、名詞以外のテストデータを使用した実験をしていくとともに、慣用句的語義の分散表現の作成方法の検討や、サポートベクターマシンを使った実験などを行い、高い精度を持つ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

[1] 佐々木 稔, 古宮 嘉那子, 新納 浩幸. 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, 言語処理学会年次大会発表論文集, P11-1 (2016).

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.