単義語の分散表現と単語間の係り受け関係を用いた語義曖昧性解消

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

自然言語に含まれている多義語を分析することは、自然言語処理の分野にとって大きな課題となっている。例えば、「研究には”意味”がある」という文の訳を計算機に判別させる場合、“意味”は「（辞書的な言葉の）意味」や「（行為の）価値」といった語義がある多義語であるため、どの語義が正しいのかを判別するための処理（語義曖昧性解消（Word Sense Disambiguation, WSD））を行わなければならない。

多義語の WSD を行う際、一般的な手法は周辺の共起単語を特徴として利用する。また、近年では word2vec を用いた単語の分散表現の研究が数多く行われており、周辺の共起単語を word2vec の分散表現へと置き換えたものを使用した WSD は有用な結果を得られることが分かっている[1]。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してしまう問題がある。例えば、岩波国語辞典の“ある”には３つの語義が存在するが、入力データでは区別がついていないので“ある”のベクトルは一つしか生成されない。この場合「ある意味」と「意味（が）ある」という二つの文の文脈を表すベクトルは同一になってしまい、二つの文で登場している「意味」という多義語の語義判別を行いたい場合に有効な特徴が得ることができない。他にも、既存の語義曖昧性解消の手法では、周辺の共起単語すべてを特徴とするため語義識別に有効ではない単語も特徴として含めてしまい精度が低下するという問題がある。

この解決方法として、本研究では単義語の分散表現と単語の係り受け関係に注目した。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。また、単語の係り受け関係を用いることで対象単語と意味的なつながりのある単語を特徴として捉えることができる。そこで本研究では、単義語の分散表現と係り受け関係にある単語の分散表現を語義曖昧性解消の素性として用い、その有効性を明らかにする。

2 関連研究

語義曖昧性解消では一般的に、多義語と共起する単語が語義識別の有効な手掛かりになるとされている[2]。本研究では、単語の分散表現、多義語と共起する単義語、多義語と係り受け関係にある単語を素性として用いる。

単語の分散表現は、単語の意味的特徴を捉える上で有効であるため自然言語処理の様々なタスクでの性能向上が期待されている。特に Mikolov ら[3]の提案した skip-gram モデルと CBoW モデルは、良質な分散表現の学習を高速に行うことができるため、近年では word2vec を利用した研究が数多く行われている[1][4]。また、菅原ら[5]の研究では、教師あり学習の語義曖昧性解消タスクにおいて分散表現の有効性が検証されている。

Li ら[6]の報告では単義語を語義曖昧性解消の素性として利用する手法を提案しその有効性が報告されている。しかし、ニュー

ラルネットワークに基づいたベクトル表現は利用しておらず、日本語の用例文での有効性については検証されていない。

また、永松ら[7]の報告では類似性判定ならびに文章検索において、周辺単語を抽出した場合よりも係り受け関係にある単語を素性として利用した場合において、よい検索結果が得られることを示している。本研究では、多義語に共起する単義語や係り受け関係にある単語の分散表現を教師あり WSD の素性として用い、その有効性の検証を行う。

3 WSD 手法

3.1 WSD モデルの学習

WSD モデルを構築するために訓練データから特徴の抽出を行い、教師あり学習手法を用いて学習を行う。この時、本研究では単義語の分散表現、係り受け関係にある単語の分散表現、及びそれらを併用した場合の特徴抽出を行った。更に、比較評価を行うために従来手法である文脈全体の分散表現による特徴抽出についても実験を行った。訓練データには対象単語の用例文に語義がついたデータを用い、あらかじめ形態素解析器で単語毎に分割する。

以下に、使用する特徴抽出の手法を列挙する。

3.1.1 文脈全体の分散表現を利用した特徴抽出（従来手法）

訓練データについて、用例文の対象単語以外の全単語を分散表現に置き換え、次元毎に加算したものを平均化することで文脈ベクトルを作成する。その後、用例文の語義と紐づけることで、それを特徴とする。

3.1.2 単義語の分散表現を利用した特徴抽出

単義語を利用するために、本研究では、岩波国語辞典から大分類に対応する語義番号の 2 番目以降がすべて 0(XXXX-0-0-0-0)かつ語義がただ一つのものだけを選び、それを単義語とした。また、それ以外の単語を多義語と定義した。学習では訓練データの用例文中の対象単語において前後 2 単語に出現する単義語を利用し、単義語一つ一つにその用例文における語義を紐づける。最後に語義の付いた単義語を分散表現へ置き換えることで、それを特徴とする。

3.1.3 係り受け関係にある単語の分散表現を利用した特徴抽出

係り受け解析器を利用して、学習モデルを作成する。まず、訓練データの用例文中の対象単語に注目し、対象単語が係る単語と対象単語に係る単語のみを抽出し、それぞれを分散表現に置き換える。次に、抽出した単語の分散表現を次元毎に加算し、平均ベクトルを取ることで対象の文の文脈ベクトルとする。作成した文脈ベクトルにその用例文の語義を紐づけることで、それを特徴とする。

3.1.4 単義語の分散表現と係り受け関係にある単語の分散表現を併用した特徴抽出

対象単語の品詞によって上記二つのどちらの手法を利用するかを区別する。動詞・形容詞・名詞（副詞可能、形容動詞語幹）なら 3.1.2、名詞（それ以外）なら 3.1.3 の特徴を用いる。

3.2 対象単語の語義推定

3.1 で説明した各特徴抽出手法から学習モデルを構築し、対象単語の語義推定に利用する。テストデータの用例文を入力することによって、対象単語の語義推定を行う。テストデータは、対象単語の用例文が含まれ、形態素解析器で単語毎に分割されている。

3.2.1 文脈全体の分散表現を利用した WSD

（従来手法）

テストデータ中の対象単語以外のすべての単語を分散表現に置き換え、次元毎に加算したものを平均化することで文脈ベクトルを作成する。その後、訓練データ全てのベクトルと類似度を比較し、最も類似しているベクトルの語義を正解として出力する。

3.2.2 単義語の分散表現を利用した WSD

訓練データと同様に、テストデータについても対象単語の前後 2 単語の単義語を利用する。抽出した単義語は分散表現に置き換えた後にそれぞれ訓練データの全ベクトルと比較し、テストデータの単義語と最も高い類似度を持つ訓練データの単義語の語義を正解として出力する。また、前後 2 単語に単義語が存在しない場合には、前後 2 単語に存在する全ての単語をテストデータとして利用する。その場合には文脈ベクトルを作成し、訓練データの中で最も類似しているベクトルの語義を正解として出力する。

3.2.3 係り受け関係にある単語の分散表現を利用した WSD

訓練データと同様に、テストデータについても対象単語と係り受け関係にある単語を利用する。係り受け関係にある単語を分散表現に置き換えた後、次元毎に加算し、平均ベクトルを取ることで文脈ベクトルとする。その後、テストデータの文脈ベクトルと最も類似度の高い訓練データの文脈ベクトルの語義を正解として出力する。また、テストデータの用例文中に対象単語と係り受け関係にある単語が存在しない場合は、前後 2 単語に存在するすべての単語をテストデータとして利用する。その場合においても、テストデータの文脈ベクトルと最も類似度の高い訓練データの文脈ベクトルの語義を正解として出力する。

表 1 ：各 WSD 手法の全体精度

使用ベクトル	従来手法(3.2.1)	提案手法(3.2.4)
朝日新聞skipgram	69.52%	70.04%
〃 cbow	69.16%	69.96%
〃 glove	69.20%	70.60%
nwjc2vec	70.16%	72.08%

3.2.4 単義語の分散表現と係り受け関係にある単語の分散表現を併用した WSD

対象単語の品詞によって上記二つのどちらの手法を利用するかを区別する。動詞・形容詞・名詞（副詞可能、形容動詞語幹）なら 3.2.2、名詞（それ以外）なら 3.2.3 の訓練データを用いた WSD 手法を行う。

4 実験

4.1 データセット

本実験では、辞書として岩波国語辞典を利用し、形態素解析器としてMeCab*1、係り受け解析器としてCabocha*2を利用した。テストデータおよび訓練データとしては、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[8]。Semeval2010 では対象単語が50個用意され、訓練データとテストデータとしてその単語を使用した用例文が各50件用意されている。

また、分散表現のデータとして、国立国語研究所が作成した nwcj2vec[9]、及び、朝日新聞のデータセットから作成したベクトル三つを利用した[10]。

4.2 結果

3 節で説明した各手法を用いて比較実験を行った。各 WSD 手法の精度を表 1 に示す。従来手法と比較すると、前後 2 単語の単義語の分散表現を使用した WSD 手法は精度が低下している一方で、係り受け関係にある単語の分散表現を使用した WSD 手法と二つを併用した WSD 手法はそれぞれ

表 2 ：3.2.4 手法の使用ベクトル毎の比較

*1 http://taku910.github.io/mecab/

*2 https://taku910.github.io/cabocha/

使用手法	全体精度
従来手法(3.2.1)	70.16%
単義語(3.2.2)	68.40%
係り受け(3.2.3)	70.56%
単義語・係り受け(3.2.4)	72.08%

精度が向上していることが分かる。特に、二つの手法を合わせた手法は従来手法と比べて 1.92%の精度上昇がみられた。

また、別のデータセットでも同様の結果を得られるか比較実験を行った。朝日新聞のデータセットから作成した分散表現 3 つと nwjc2vec を用いた場合の精度を表 2 に示す。従来手法と比べるといずれの場合も精度が 0.52%~1.92%上昇していることが分かる。

以上のことから、対象単語の周辺に出現する単義語は主に名詞の語義識別に有効な特徴を持っていることが分かる。その一方で、対象単語の品詞が、副詞的な用法や形容動詞的な用法で用いられる名詞、動詞、形容詞の場合、共起する単義語は有効な特徴でないことも分かった。また、係り受け関係にある単語は対象単語の品詞に関わらず有効な特徴を持っていることも分かった。このように対象単語の品詞毎に特徴を捉えた WSD 手法を選択することで、全体の精度を上昇させることができると考えられる。

5 結論

本研究では語義曖昧性解消タスクにおいて、対象単語周辺に出現する単義語の分散表現と対象単語と係り受け関係にある単語の分散表現を利用したシステムの有効性を検証した。実験の結果、二つの手法を併用したシステムが従来手法の精度を上回ることが確認され、これらの手法が有効であることが明らかになった。今回は MeCab における品詞区分により、単義語の分散表現を利用した手法と、係り受け関係にある単語の分散表現を利用した手法を選択したが、品詞毎の特徴をより捉えた手法を利用することでさらなる WSD システムの性能向上に寄与することが期待できる。

参考文献

[1] 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.

[2] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In Proc. of ACL 1995, pages 189–196, 1995.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26,

[4] X. Chen, Z. Liu, and M. Sun. A unified model for word sense representation and disambiguation. In Proc. of EMNLP 2014, pages 1025–1035, 2014.

[5] 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 発表論文集 (2015 年 3 月)pp.648-651

[6] J. Li, and C. Huang(1999). “A Model for Word Sense Disambiguation” Cmputational Linguistics and Chinese

[7] 永松 健司・田中 英彦「コーパスから抽出した係り受け共起情報に基づく類似度と文書検索における評価」自然言語処理(1996)116-11,pp.73-78

[8] Okumura, M., Shirai, K., Komiya, K., and Yokono, H. (2011). “On SemEval-2010 Japanese WSD Task.” 自然言語処理, 18 (3), pp. 293–307.

[9] 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.

[10] 田口雄哉, 田森秀明, 人見雄太, 西鳥羽二郎, 菊田洸. 同義語を考慮した日本語単語分散表現の学習. 情報処理学会第 233 回自然言語処理研究会, Vol.2017-NL-233, No.17, pp.1-5. October 2017