単語の類似性を考慮した括弧表現分類の精度に関する考察

村田 浩章　　指導教員：山村 毅

1　はじめに

我々が普段使用する文章には，丸括弧や鉤括弧などの括弧表現が多く使われる．文章中の括弧表現には様々な用法が存在し，同じ丸括弧や鍵括弧でも複数の用法で使用されることから，括弧表現にはそれぞれ意味があるといえる．そのため，文章に対して構文解析等の機械処理を行う際にも，括弧表現の扱いを意味ごとに変える必要がある．本研究では，括弧表現の意味を分類問題の枠組みで考える．

自然言語処理における分類問題では，まずはじめに，文を特徴表現する方法を考える必要がある．最もよく知られる手法の一つに bag of words がある．これは，文章中の単語の出現回数をそのまま特徴とする手法である．しかしながら，bag of words では類似した単語に対応できないという問題点がある．

本研究では，単語の類似性を考慮するために word2vec を用いる．特徴として，各単語をそのまま特徴とした場合（Bag of words），word2vec によって単語をベクトル化したものを特徴とした場合 (word embedding)，ベクトル化した結果を用いてクラスタリングしたものを特徴とした場合 (Clustered bag of words) の 3 つの手法で括弧表現の分類実験を行う．また，ベクトル化に際して各単語を原形で使用するか表層形で使用するか，品詞情報ありか品詞情報なしかの組み合わせ 4 通りを試すことで，結果にどのような違いが現れるかを考察する．

2　先行研究

先行研究 [1] では，本文に対してその括弧記号が必要か不必要か，またその処理方法の違いに基づいて「補足情報」,「引用」,「独立文」,「小見出し」,「補完」,「記号表現」の 6 つの分類カテゴリを設けた．そして，2005 年の毎日新聞の 5 日分の記事から括弧表現を 7809 個抽出し人手でラベル付けした．各カテゴリのデータ数を表 1 に示す (表中の補足は補足情報，独立は独立文，小見は小見出し，記号は記号表現を表す)．

表 1　各カテゴリのデータ数

ラベル	補足	引用	独立	小見	補完	記号
数	3384	2915	438	585	103	388

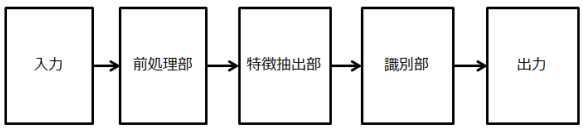

図 1　括弧表現分類のシステム構成

括弧表現分類は，図 1 の手順に従って行った．すなわち，まず，入力として得られた括弧表現を含む文章データに対し，前処理として形態素解析を行う．次に，文章から特徴を抽出し，SVM によって分類を行い，最後に分類結果を出力する．実験の結果，全体として 95.1% の F 値が得られた．

3　類似性を考慮した特徴

括弧表現分類の精度向上には，特徴抽出部における特徴の選択，もしくは識別部による分類器の選択が必要になると考えられる．そこで，本研究では，特に特徴抽出部における特徴選択に注目することにする．

3.1　Bag of words の問題点

文章の特徴表現に使われるものとして，最もよく知られる手法の一つに Bag of words がある．これは，文章に単語が含まれているか否かを特徴とするものであるが，Bag of words では，類似した単語に対応できていないという問題点がある．

自然言語処理における分類問題では，類似した文や単語についての結果を利用するのが一般的である．これは，全く同じ文書についての分類データを持っているのが理想ではあるが，一般的にはそのような場合は極めて少ないからである．

従って，Bag of words を用いて単純に単語の一致率を見るだけでは，特に学習データの数が少ない場合，十分な性能が得られない可能性がある．

3.2　Word Embedding

一般によく使われる Bag of words では，単語を単なる識別子として捉えるため単語間の類似性を直接扱えなかった．そこで単語を数値の分散表現（ベクトル）で表現するというアイデアが現れた．これを Word Embedding という．

Word Embedding の手法の一つに word2vec[2] がある．これは，2013 年に Tomas Mikolov が提案した，単語をベクトル化して表現するための手法である．近い意味を持つ単語は似たような文脈で使われている，という考えのもと単語をベクトル表現することで，似た意味を持つ単語はベクトル空間上で近くに配置される．これを用いることで，単語や文章間の類似度を計算することが可能になる．

3.3　Clustered bag of words

word2vec を用いて単語をベクトル化したものを特徴とすることで，単語の類似度を考慮することにより，Bag of words と違い，少ないデータでも一般性を獲得できると考えられる．しかしながら，単語をベクトル化する際，ベクトルの次元数が大きいほど，似た単語であっても距離が離れていき，各単語の距離の差は小さくなるという問題がある．

そこで，各単語をベクトル化したものをクラスタリングし，単語がどのクラスに属しているかを特徴とすることで，類義語としてのまとまりを明確にし，単語の意味を考慮した Bag of words として，より一般性を得ることが出来ると考えられる．

4　括弧表現分類のための特徴

本研究では，括弧表現分類において，括弧に近い単語だけに注目すれば分類できると考え，括弧の前後と中の単語のみを特徴とする．また，括弧の種類については，重要な特徴の一つだと考えられるため，これも特徴に加える．単語を特徴とするとき，3 で述べた問題に対処するため，次の 2 つの方法で特徴表現する．

(1) word embedding
括弧の直前，中の最後，直後の単語に注目して，各単語を word2vec を用いて求めたベクトル表現に置き換えたものを特徴とする．

(2) Clustered bag of words
括弧の直前，中，直後の単語クラスに注目して，単語が含まれるクラスに対応するベクトルを 1，含まれていないベクトルを 0 としたものを特徴とする．単語クラスを求めるのに，word2vec を用いて求めた単語ベクトルをクラスタリングする．

5 実験と結果

実験データは先行研究と同じものを用いる．特徴は，4 で述べた方法に加え，比較のために Bag of words を特徴としたものを使用して実験を行う．このとき，特徴表現をするにあたり，文章中の単語を原形に直すか，表層形のまま使用するか，品詞情報を加えるかどうかの違いによる精度がどのように異なるかを調べた（以下，原形品詞情報無しを O−，原形品詞情報有りを O+，表層形品詞情報無しを S−，表層形品情報有りを S+ とする）．形態素解析には MeCab を使用し，ツールとして Weka，word2vec を使用する．word2vec の学習には 2005 年の毎日新聞一年分のデータを使用した．また，学習モデルは Skip-gram を使用し，ベクトルの次元数は 300，クラスタリングは k 平均法を用いて 400 のクラスに分けた．分類器には SVM を使用し，10 分割交差検定を用いて実験を行った．

各手法における S− の F 値についてまとめたものを表 2 に示す (表中の Bow は Bag of words，we は word embedding，C-bow は Clustered bag of words を表す)．また，各手法毎に，原形，表層形，品詞情報の違いによる F 値についてまとめたものを表 3 から表 5 に示す（表中の補足は補足情報，独立は独立文，小見は小見出し，記号は記号表現を表す）．

表 2 各手法における S− の F 値

ラベル	補足	引用	独立	小見	補完	記号	全体	平均
Bow	.978	.990	.939	.935	.809	.926	.972	.929
we	.975	.990	.923	.929	.888	.914	.970	.936
C-bow	.983	.990	.925	.945	.919	.917	.975	.946

表 2 をみると，特にデータ数の少なかった補完カテゴリにおいて，word embedding ，Clustered bag of words は，Bag of words と比べて F 値が高い．これは，単語の類似性を考慮することで，データの少なかったカテゴリについても一般性を獲得することができたため，分類精度が向上したと思われる．次に，小見出しカテゴリをみると，Clustered bag of words は他の 2 つの手法と比較して F 値が高い．このことから，単純にベクトル化した結果を使うだけではなく，クラスタリングした結果を使用することで，分類精度が向上する可能性があることがわかる．

また，表 3 を見ると，Bag of words では，全てのカテゴリで原形・表層形・品詞の有無による差は最大でも 0.5pt であり，F 値に差はほとんどなかった．さらに，表 4 をみると，word embedding においても，同様に単語の形や品詞情報ではほとんど差はないといえるが，O+，S+ の分類結果は O−，S− に比べ低くなる傾向にある．一方，表 5 をみると，Clustered bag of words では，S− における補完カテゴリと O− における記号カテゴリが，他の単語の種類と比較して高い F 値を出している．また，word embedding と同様に，O+，S+ の分類結果は O−，S− に比べ低い傾向にある．今回の実験では，学習の際に使う単語の種類に関しては，実験結果に与える影響は小さいものの，品詞情報を加えることにより F 値が下がる傾向にあることがわかる．

6 まとめ

本研究では，単語の類似性を考慮するために word2vec を使用し，3 つの手法で分類実験を行った．また，各手法において単語の原形，表層形の違いと品詞情報の有無によって精度がどのように異なるかを調べた．

その結果，単語の類似性を考慮することで，データの少ないカテゴリについても一般性を獲得することができた．さらに，ベクトル化した単語をクラスタリングすることによる，精度向上の可能性を示した．また，単語の扱い方の違いでは F 値に差ほとんどでなかったが，品詞情報は括弧表現分類においてノイズになっている可能性があることがわかった．

今後の課題としては、括弧の直前，直後だけでなく，より多くの単語に注目して特徴抽出することや，ベクトル化した単語のクラスタリングに，k 平均法以外の方法を用いることなどが挙げられる．

参考文献

[1] 村田浩章，落合里咲，山村毅：最大エントロピーモデルを用いた括弧表現分類，電気・電子情報関係学会東海支部連合大会講演論文集，L4-2，2014

[2] Tomas Mikolov，Wen-tau Yih，Geoffrey Zweig： “Linguistic Regularities in Continuous Space Word Representations”，NAACL-HLT，2013.

表 3 Bag of words

ラベル	O−	O+	S−	S+
補足	.978	.978	.978	.978
引用	.990	.989	.990	.989
独立	.937	.938	.939	.937
小見	.933	.934	.935	.936
補完	.809	.804	.809	.809
記号	.927	.927	.926	.926
全体	.972	.972	.972	.972
平均	.929	.928	.929	.929

表 4 word embedding

ラベル	O−	O+	S−	S+
補足	.975	.975	.975	.975
引用	.989	.988	.990	.988
独立	.916	.910	.923	.915
小見	.936	.935	.929	.928
補完	.887	.890	.888	.897
記号	.905	.909	.914	.904
全体	.969	.969	.970	.968
平均	.935	.934	.936	.934

表 5 Clustered bag of words

ラベル	O−	O+	S−	S+
補足	.980	.978	.983	.978
引用	.990	.990	.990	.990
独立	.926	.931	.925	.925
小見	.949	.944	.945	.943
補完	.838	.863	.919	.903
記号	.932	.892	.917	.900
全体	.974	.972	.975	.972
平均	.936	.933	.946	.940