係り受け関係を用いた 短単位の単語ベクトルから長単位の単語ベクトルの合成

清藤 拓実　古宮 嘉那子　佐々木 稔　新納 浩幸
茨城大学工学部情報工学科

{14t4037r,
kanako.komiya.nlp,minoru.sasaki.01,hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

1　はじめに

近年、word2vec をはじめとする分散表現を利用した研究が盛んである。一方、単語境界が曖昧である日本語では、複数の言語単位が存在している。例えば unidic の短単位では「会員」は一単語であるが、「裁判員」は二単語であり、これらを直接比較することができない。そのため、複合語を単語から合成する手法が必要である。本研究では unidic で利用されている単語単位の「短単位」の単語ベクトルから「長単位」の単語ベクトルを合成する。この際、複数の単語の分散表現から句の分散表現を合成する研究 [1] を参考に、長単位の単語ベクトル内の係り受け関係を考慮した合成を行った。

2　関連研究

近年、複数の単語の分散表現から句の分散表現の合成を行う研究が注目されている [1][2][3][4]。本研究では過去の句の分散表現を生成する手法を応用し、長単位の分散表現を生成する。

Lexfunc[2] のモデルは係り受けを用いたモデルではあるが、単語の表現の形式が従属語は行列で、独立語はベクトルとなっており統一されていない。

村岡ら [1] は、Lexfunc[2] モデルに再帰性がないことや、RNN モデルでは明らか修飾関係の異なる句 (例: 形容詞＋名詞と動詞＋名詞など) を同じものとして学習を行っていることを問題視し、修飾関係を考慮した新たなニューラルネットワークモデルを提唱した。

橋本ら [3] は、テンソル分解を用いて熟語と項の関係を掛け算的にモデル化し、動詞句の表現を効率的に学習する手法を提唱した。

また橋本ら [4] は構成性と非構成性を同時に学習する手法を提唱した。橋本ら [4] は、句の意味はその構成要素の単語の意味の組み合わせによって決まると仮定している。同様に本研究では、長単位は四字熟語など特殊な例を除き、短単位の意味の合成によって成り立っていると仮定する。例として

茨城大学工学部

について考えてみる。「茨城大学工学部」という長単位はそれぞれ「茨城」＋「大学」＋「工学部」という短単位に分けられる。また茨城県にある大学の工学部であるというように、これらの短単位の組合せによって長単位の単語の意味を推測することができる。このように長単位は、四字熟語などの特殊な例外を除き、構成される短単位の意味から類推することができる。そのため短単位の単語の分散表現から長単位の分散表現を合成できるのではないかと考えた。

3　提案手法

本研究は村岡ら [1] の係り受け関係を用いた句ベクトルの生成をベースにした。村岡ら [1] は、単語ベクトルを入力とし、句ベクトルの合成を行っている。村岡らの手法では入力を 2 つの単語ベクトル u,v、出力を句ベクトル p としているが、我々は入力が短単位の単語ベクトル、出力を長単位の単語ベクトルなので、二つの短単位の単語ベクトルを s_1,s_2、長単位の単語ベクトルを l として定式化する。係り受けを考慮しないで行う長単位を短単位から生成するときの入力は短単位の分散表現のみである。これを一般的な式で表す

と次式で表される。

$$l = f(s_1, s_2) = \sigma \left(W \begin{bmatrix} s_1 \\ s_2 \\ b \end{bmatrix} \right) \quad (1)$$

この式は s_1、s_2 の 2 つの短単位を入力した際の計算を示している。ただし、s_1、s_2 は d 次元のベクトルを表し、W は $d \times (2d+1)$ 行列であり、b はバイアスである。また σ はシグモイド関数を意味する。しかし、この式では、「発表会」(「会」の説明を「発表」で行う) や「12 時」(数詞＋単位) のように明らかに異なる性質の組み合わせを同じ重みで表すことになる。ここで本研究では [1] と同様に、異なる性質を別の重みとして分けるために入力として係り受け関係 r を加えた。

$$l = f(s_1, s_2, r) = \sigma \left(W_r \begin{bmatrix} s_{1r} \\ s_{2r} \\ b_r \end{bmatrix} \right) \quad (2)$$

そのため、ニューラルネットの重みとバイアスが係り受けによって変更されることになる。ここで W_r と b_r は係り受けの数 (本実験では 13 通り) 用意する。本実験での誤差関数は村岡ら [1] の手法で用いていた誤差関数と同様に以下の式を用いた。

$$j(\theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} ||p_i - t_i||^2 + \frac{\lambda}{2} ||\ \theta\ ||^2 \quad (3)$$

ただし θ は学習パラメータの全て、p_i は生成したベクトルの集合、t_i は教師データの集合を表す。勾配の計算は誤差逆伝播を用いた。

4 実験

本節では、初めに入力および教師ベクトルとなる短単位・長単位の単語の分散表現の作成方法を説明し、次に係り受けの定義を明示し、評価方法を明らかにしたうえで、最後に実験の結果を記載する。

4.1 短単位・長単位の分散表現の作成

短単位・長単位の分散表現は以下のような手順で作成する。

1. 『現代日本語書き均衡コーパス』[5] の分かち書きを行う。この時、短単位の単語の分かち書きに長単位の単語の分かち書きを追記したファイルを作成する。
2. 分かち書きのファイルを用いて分散表現を生成する。

分散表現の作成には word2vec [1] を利用した。この際、skip-gram を使用し、ベクトルの次元数は 100 次元とし、他のパラメータはデフォルトの値を用いた。用いる長単位の定義を「長単位を構成する短単位の数が 2 つのもの」としているため、短単位と長単位が同じものと長単位を構成する短単位の数が 3 単語以上のものは除外した。

以上のようにして計 169,736 単語の長単位の分散表現とそれらを構成する短単位を取得した。

4.2 係り受けの定義

長単位の単語の係り受けは日本語の熟語の構成をもとに定義した。まず、『現代日本語書き均衡コーパス』の書籍全般のコーパス中の 23,000 件の長単位の単語の係り受けを人手で分類し、それらを訓練データとしてサポートベクターマシン（SVM）を用いてコーパス全体に係り受けを付与した。また SVM で liblinear [2] を用い、カーネルは線形カーネルを使用した。なお、人手で分類する際には、一つの係り受けのサンプル数が 30 件以上存在するように留意した。本研究で利用した係り受けは以下の 13 通りである。

1. 前の短単位が後の短単位の説明を行う組合せ
   例:「講習会」
2. 目的語と述語の組合せ
   例:「債務放棄」
3. 補語と述語の組合せ
   例:「法的整理」
4. 主語と述語の組合せ
   例:「画面割れ」
5. 一方の短単位がもう一方の短単位の単位となる組合せ
   例:「1 月」
6. 主要単語と接尾語の組合せ
   例:「具体的」
7. 接頭語と主要単語の組合せ
   例:「副代表」
8. 片方の短単位に、助詞が用いられている組合せ
   例:「ための」
9. 固有名詞と一般名詞の組合せ
   例:「茨城県」

[1] https://radimrehurek.com/gensim/
[2] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

10. 名詞と動詞で動詞になる組合せ
    例:「応募する」
11. 数字どうしの組合せ
    例:「三二」
12. 短単位単体では意味を持たず長単位になって初めて意味を持つ組合せ
    例:「だが」
13. その他
    例:「意気揚々」

4.3 ネットワークモデル

本研究ではフィードフォワードニューラルネットワーク（FFNN）を用いて長単位の単語ベクトルを合成した。入力は 2 つの短単位の単語ベクトル、出力は長単位の単語ベクトルとした。中間層のユニット数は 300 とした。エポック数は最大を 2,000 とした。重みの初期値はランダムに生成した。

4.4 評価方法

取得した分散表現から半数を訓練データ、もう半数をテストデータとして二分割交差検定を行った。この際、生成した単語の長単位の分散表現とモデルにより生成した分散表現とのコサイン類似度により評価した。コサイン類似度が高ければ高いほど、本来取得したい分散表現と近いということになるため、性能が高いモデルであるといえる。

5 結果

表 1 は係り受け無しでモデルを学習させた場合のコサイン類似度と、係り受けごとに重みを分けた場合のコサイン類似度を比較したものである。

表 1: 係り受けの有無のコサイン類似度の比較

エポック数	係り受け無し	係り受けあり
200	0.283	0.415
400	0.328	0.500
600	0.387	0.548
800	0.425	0.575
1,000	0.468	0.96
1,200	0.491	0.607
1,400	0.515	0.615
1,600	0.528	0.621
1,800	0.545	0.626
2,000	0.555	0.629

表 1 を見ると係り受けありのほうが早い段階で目標データとの類似度の高い結果を得ることができていることがわかる。

次に表 2 は係り受け毎のコサイン類似度を列挙したものである。この時エポック数は 200 から 2,000 まで 200 おきに結果を出しその中で最も性能が良かった結果を示した。表 2 のエポック数とは表の値を出力した際のエポック数である。表 2 から名詞と動詞で動詞になる組合せである「係り受け 10」が最も性能が良く、短単位単体では意味を持たず長単位になって初めて意味を持つ組合せである「係り受け 12」が最も性能が低くなることが分かった。

表 2: 係り受け毎のコサイン類似度の比較

係り受け	コサイン類似度	エポック数
1	0.608	2,000
2	0.656	1,600
3	0.61	600
4	0.668	800
5	0.617	2,000
6	0.630	2,000
7	0.619	2,000
8	0.62	400
9	0.619	2,000
10	0.722	2,000
11	0.666	600
12	0.470	200
13	0.632	2,000

6 考察

本実験の結果により、係り受け関係を考慮して短単位の分散表現の合成を行ったほうが学習の性能向上につながることが確認できた。

本研究では簡略化のため、単語の制限として 2 つの短単位からなる長単位に範囲に絞って実験を行った。しかし、入力に用いた単語情報と出力に用いた単語情報はともにベクトルであるため、3 単語以上の合成を行うことが可能である。

本手法では係り受けの付与を行う際の SVM の分類の正答率がモデルの性能に影響する。そこで性能向上を図るため分類の正答率を評価した。それぞれの係り受けに対してランダムにサンプルを 100 ずつ取得し、そのサンプルの分類を人手で正誤を判定した。実験結果を表 3 に示す。

表 3 から正答率が他に比べて補語と述語の組合せである「係り受け 3」と、主語と述語の組合せである「係り受け 4」の分類の精度が極めて低いことがわかる。

表 3: 係り受け分類の最も多い誤りと正答率

係り受け	最も多い誤り	正答率 (%)
1	6	84
2	1	42
3	1	8
4	1	16
5	6	99
6	13	91
7	1,6,13	94
8	1,6	85
9	1	91
10	1	99
11	1	86
12	1	86
13	1	84

そこで「係り受け 3」と「係り受け 4」を誤りの多かった、前の短単位が後の短単位の説明を行う組合せである「係り受け 1 」と統合し、「係り受け 1'」として再実験を行う。再実験は表 4 で示す。

表 4: 係り受け 1' を用いた実験

エポック数	全体のコサイン類似度	係り受け 1'
200	0.402	0.327
400	0.489	0.415
600	0.541	0.480
800	0.571	0.521
1,000	0.590	0.547
1,200	0.603	0.567
1,400	0.613	0.583
1,600	0.620	0.594
1,800	0.624	0.602
2,000	0.629	0.609

表 1 と表 4 を比較するとわずかに合成の性能が落ちたことが確認できる。原因としては、「係り受け 1」はサンプルが多いデータであり、「係り受け 3」・「係り受け 4」は「係り受け 1」よりも性能がよいデータである。したがって、「係り受け 1」に飲み込まれる形となり、性能が落ちてしまったと考えられる。

係り受けごとの精度比較を見ると、短単位単体では意味を持たず長単位になって初めて意味を持つ組合せである「係り受け 12」が著しく精度が低いことが確認できたため、助詞などの扱いを工夫することにより性能の向上する可能性がある。

7　おわりに

本稿では、短単位の分散表現から長単位の分散表現を合成する手法を提案した。本研究では係り受け関係を用いるか用いないかによって、少ないエポック数で性能の良いモデル作成をできることを確認した。

謝辞

本研究は，茨城大学の女性エンパワーメント支援制度補助金および JSPS 科研費 15K16046 の助成を受けたものである。

参考文献

[1] 村岡雅康, 島岡聖秀, 山本風人, 渡邉陽太郎, 岡崎直観, 乾健太郎, 係り受け関係を用いた句ベクトルの生成　言語処理学会　発表論文集 pp.1055-1058, 2014.

[2] M. Baroni and R. Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In EMNLP, pp. 1183-1193, 2010.

[3] 橋本和真, 鶴岡慶雅 テンソル分解に基づく述語項構造のモデル化と動詞句の表現ベクトルの学習 言語処理学会　発表論文集 pp.639-642, 2015.

[4] 橋本和真 鶴岡慶雅 構成性と非構成性を同時に考慮した動詞句の表現学習 言語処理学会　発表論文集　pp.661-664, 2016.

[5] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, Yasuharu Den, Balanced Corpus of Contemporary Written Japanese,Language Resources and Evaluation, Vol.48, pp.345-371, 2014.