Vol.2020-NL-246 No.23
Vol.2020-SLP-134 No.23
2020/12/3

品詞情報を利用した複合語の分散表現の合成

河野 慎司[1,a)]　古宮 嘉那子[1,b)]

概要：本稿では, 複合語の分散表現をその構成語から合成する試みについて述べる. 具体的には UniDic での単語区切りの単位である「短単位」と「長単位」をそれぞれ構成語と複合語として利用し, 構成語二つの分散表現から複合語の分散表現の合成をニューラルネットワークで行った. 複合語の分散表現の教師値は, 構成語で分かち書きしたコーパスと複合語で分かち書きしたコーパスを連結して分散表現を作成した. 分散表現の合成にあたって, 現代日本語書き言葉均衡コーパスに付与されている品詞パターンを分類するタスクをサブタスクとして利用してマルチタスク学習を行った. サブタスクを利用した場合としない場合とで, 合成した複合語の分散表現と正解の分散表現のコサイン類似度を比較した結果, サブタスクを利用した方が, 合成性能が高いことが分かった. また構成語間の意味的関係を使った先行研究と比較したところ, 本手法の性能が上回った.

Composing Word Embeddings for Compound Words Using Patterns of Parts of Speech

Abstract: This paper describes an attempt to compose word embeddings of a compound word from its constituent words. In particular, we used “short unit” and “long unit” which are the units of word delimiter in UniDic for constituent and compound words respectively, and compose a word embedding of a compound word from word embeddings of two constituent words using a neural network. The supervised data of the word embedding of compound words was created with a corpus generated by concatenating the corpus divided by the constituent words and the corpus divided by the compound words. In the composition of word embedding, multitask learning was performed using the task of classifying the parts of speech patterns assigned to “Balanced Corpus of Contemporary Written Japanese” as a subtask. We compared the cosine similarity between the composed and correct word embeddings of compound words to assess the models with and without the subtask. The experiments revealed that the model with the subtask outperformed the model without the subtask. In addition, the performance of this method was superior to that of a previous study using semantic information.

1. はじめに

近年, 単語の分散表現は自然言語処理における基幹技術となっている. 分散表現を取得するにあたって, まず単語を分かち書きをする必要がある. しかし単語境界の違いによっては単語の分散表現を取得後, 単語同士を直接比較ができないときがある. 例えば Unidic の短単位では「会員」は一単語であるが, 「裁判員」は二単語であるため直接比較ができない. そのため複合語の分散表現を, その構成語の単語の分散表現から合成する手法が必要である. またコーパス中に存在する単語を組み合わせて合成することで, コーパス中に存在しない単語の分散表現を生成することができる.

本研究では Unidic で利用されている単語区切り単位の「長単位」と「短単位」をそれぞれ複合語とその構成語とみなし，短単位の単語の分散表現から長単位の単語の分散表現の合成を行う.

河野ら [6] は人手で意味的関係を定義したデータを用意し, その分類タスクを用いたマルチタスク学習を利用して, 短単位の単語から長単位の単語の合成を行っている. しかしこの手法は，人手で短単位の単語同士の意味的関係を付与する必要があるため, 分類タスクのデータ数が多くないことや, 正しい意味的関係が付与できているか分からないといった問題がある.

[1] 茨城大学大学院理工学研究科情報工学専攻
Ibaraki University, Nakanarusawa 4-12-1, Hiachi, Ibaraki 316–8511, Japan
[a)] 20nm709n@vc.ibaraki.ac.jp
[b)] kanako.komiya.nlp@vc.ibaraki.ac.jp

本研究では, これらの問題を『現代日本語書き言葉均衡コーパス (以下, BCCWJ)』の品詞パターンを利用することで解決を試みた. 具体的には，まず, 事前学習として短単位 2 つから長単位 1 つを合成するタスクと, 構成語間の関係を分類するタスクを同時学習するマルチタスク学習モデルを構築する. このとき, 構成語間の関係として，BCCWJ から取得できる品詞情報のパターンを用いた．本稿では，マルチタスク学習を行わない基本的なモデルや，先行研究との比較を行う.

2. 関連研究

本研究の関連研究として, 句の分散表現の生成の研究が挙げられる. Muraoka ら [4] は「形容詞+名詞」や「名詞+名詞」などの係り受け関係に注目し, それぞれの係り受け関係毎に異なる重みで分散表現の合成を行っている. Hashimoto ら [1] は句の表現は構成要素の単語の表現から計算できるという「構成的表現」とイディオムとして扱われる「非構成的表現」の両方を同時学習する手法を提案している. ここで Hashimoto ら [1] 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.

コーパスに存在しない単語の分散表現を生成する研究には，[2]，[3]，[5]，[6]，[7] が挙げられる．Yuval ら [5] はコーパスに存在しない単語の分散表現を取得するために, MIMICK という手法を提案している. MIMICK は文字単位の分散表現を双方向 LSTM に入力し, それらの文字から成る単語分散表現を合成する手法である.

Komiya ら [3] は，本研究と同様, 単語の比較を行う対象語同士が, 単語境界によって単語数が違うため比較が難しいという課題を解決するため，Unidic の短単位と長単位を利用して，構成語から複合語の分散表現を作成している. この研究は, 構成語間の意味的関係を 13 種類に定義し, SVM を使用して約 17 万件の複合語を意味的関係ごとに分類したのち, それぞれの意味関係ごとに分散表現を合成しており, 提案手法の方が, 全データを使用して一括して合成するよりも, 少ないエポックで性能のよい分散表現を作成できることを示した. 河野ら [6] は Komiya ら [3] を発展させ, Komiya らが SVM で事前に行っていた意味的関係の分類を分散表現の合成と同時に行う, マルチタスク学習のモデルを提案している. 本研究は, この研究を改良したものである.

Hirabayashi ら [2] は Komiya ら [3] と河野ら [6] と同じ問題を，教師なし学習を使って解決を試みている．具体的には, 短単位で取得した分散表現と, 長単位で取得した分散表現間に Bilingual Word Embeddings によるマッピングを適応し, 異なる分散表現間上での単語の分散表現の比較を行っている.

久本ら [7] は形態素解析器 Sudachi の複数粒度分割を活用し, 超大規模コーパス NWJC より取得した分散表現より, 双方向 LSTM を使用して構成語から複合語の合成を行っている.

3. 提案手法

基本となる複合語作成のためのニューラルネットワークの構成を図 1 に示す. 入力は複合語の二つの構成語（短単位）の分散表現であり, 出力は複合語（長単位）の分散表現である. モデルとしては多層パーセプトロンを使用した. この際, 複合語の分散表現の教師値には, 構成語で分かち書きしたコーパスと複合語で分かち書きしたコーパスを連結して分散表現を作成し, 利用した.

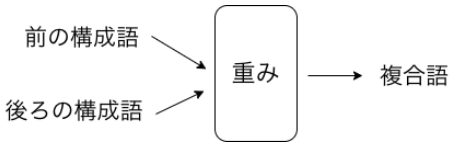

図 1 複合語の分散表現をその構成語の分散表現から合成するネットワーク

次に, 提案手法のネットワークの構成を図 2 に示す. 構成語から複合語の分散表現を合成すると同時に, 複合語の構成語の関係を分類するマルチタスク学習のモデルである. 基本のネットワークと同様, 多層パーセプトロンを利用している. 本研究の目的は, 複合語の分散表現の合成であるため, 以下, 複合語と構成語の関係を分類するタスクをサブタスクと呼ぶことにする. 河野ら [6] は, サブタスクとして, Komiya ら [3] が定義した, 構成語間の意味的関係の分類を利用している. しかし, これらの関係は, 人手でアノテーションするのに時間がかかり，データを集めるのが困難であるという問題があった. そこで, 本研究では, 自動的に取得可能な関係として, 複合語の構成語の品詞パターンに注目し, その分類をサブタスクとしたマルチタスク学習のモデルを提案する.

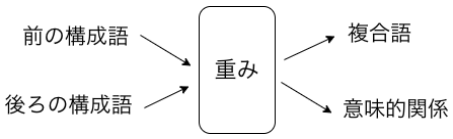

図 2 複合語と意味的関係を出力するネットワーク

マルチタスク学習の手順を以下に述べる. まず, 図 2 のようなマルチタスク学習のモデルを学習する. その重みを利用して FT を行い, 図 1 の枠組みで再学習する.

3.1 品詞パターンの定義

意味的関係には BCCWJ に付与されている品詞情報を利用する. 表 1 に例を示す.

表 1 の「警察メディア」は長単位であり, 短単位「警察」

表 1　BCCWJ に付与されている品詞情報

短単位	品詞情報
“	補助記号-括弧開
警察	名詞-普通名詞-一般
メディア	名詞-普通名詞-一般
”	補助記号-括弧閉
が	助詞-格助詞
成立	名詞-普通名詞-サ変可能
する	動詞-非自立可能

と「メディア」を構成語とした複合語としてとらえることができる．これらの構成語の品詞パターン，「名詞-普通名詞-一般＋名詞-普通名詞-一般」を正解ラベルとして，品詞パターンの分類タスクを設定した．

本実験に利用する品詞パターンには，BCCWJ から得られた短単位 2 つと長単位 1 つの全組み合わせのうち，重複を省いたデータで最も品詞パターンが多い順の 13 個 (上位 12 個+それ以外をその他とする) を採用した．以下に BCCWJ から取得した品詞パターン 13 通りを示す．かっこの中の数字は，分類タスクの訓練事例として利用した件数である．同じ複合語がコーパス中に複数回出てきたときは，「その他」以外の品詞パターンで，すべてを訓練事例として利用した．「その他」に関しては，件数が多くなりすぎるため重複を省いて利用した．

(1) 名詞-普通名詞-一般+名詞-普通名詞-一般 (11220 件)
例: ラブ+レター

(2) 名詞-普通名詞-サ変可能+動詞-非自立可能 (29770 件)
例: 発見+し

(3) 名詞-普通名詞-一般+接尾辞-名詞的-一般 (9122 件)
例: 好奇+心

(4) 名詞-普通名詞-サ変可能+名詞-普通名詞-一般 (5036 件)
例: 編集+委員

(5) 名詞-普通名詞-一般+名詞-普通名詞-サ変可能 (4511 件)
例: 自己+喪失

(6) 名詞-数詞+名詞-普通名詞-助数詞可能 (7190 件)
例: 三+月+三月

(7) 名詞-固有名詞-地名+名詞-普通名詞-一般 (3067 件)
例: 日本+語

(8) 動詞-一般+動詞-非自立可能 (967 件)
例: 遊び+慣れ

(9) 名詞-普通名詞-サ変可能+接尾辞-名詞的-一般 (5022 件)
例: 建築+家

(10) 接頭辞+名詞-普通名詞-一般 (2997 件)
例: 不+機嫌

(11) 名詞-普通名詞-サ変可能+名詞-普通名詞-サ変可能 (1610 件)
例: 再生+願望

(12) 名詞-固有名詞-人名+名詞-固有名詞-人名 (874 件)
例: 赤瀬川+原平+赤瀬川原平

(13) その他 (8545 件)
例: 初めて+づくし

以下に，比較手法のサブタスクとして利用する，河野ら [6] が使用した意味的関係の定義を示す．かっこの中の数字は，分類タスクの訓練事例として利用した件数である．なお，重複したデータは含まれていない．

(1) 前の短単位が後の短単位の説明を行う組合せ (447 件)
例:「講習会」

(2) 目的語と述語の組合せ (54 件)
例:「債務放棄」

(3) 補語と述語の組合せ (13 件)
例:「法的整理」

(4) 主語と述語の組合せ (12 件)
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 (132 件)
例:「1 月」

(6) 主要単語と接尾語の組合せ (200 件)
例:「具体的」

(7) 接頭語と主要単語の組合せ (31 件)
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 (330 件)
例:「ための」

(9) 固有名詞と一般名詞の組合せ (97 件)
例:「茨城県」

(10) 名詞と動詞で動詞になる組合せ (371 件)
例:「応募する」

(11) 数字どうしの組合せ (27 件)
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 (32 件)
例:「だが」

(13) その他 (114 件)
例:「意気揚々」

Vol.2020-NL-246 No.23
Vol.2020-SLP-134 No.23
2020/12/3

4. 実験

4.1 使用データと分散表現

本研究では, コーパスとして BCCWJ を利用する. BCCWJ は Unidic の定める, 長単位および短単位という単語の単位で分かち書きされており, この長単位を複合語, 短単位を構成語としてとらえることができる. 複合語の分散表現を合成するための訓練事例として, BCCWJ から短単位ふたつで長単位ひとつを構成している組み合わせを抜き出して利用した. これらを分散表現合成用データと呼ぶ. これらのうち, 構成語の意味的関係をアノテーションしたデータを, 比較手法のサブタスクの訓練事例として利用する. これらを意味的関係付きデータと呼ぶ. また, 分散表現合成用データのうち, 品詞パターンの付与されたデータを, 提案手法のサブタスクの訓練事例として利用する. これらを品詞パターン付きデータと呼ぶ. 以上のタスクごとの訓練事例の件数を表 2 にまとめる.

分散表現の合成の教師値として利用した分散表現は, BCCWJ を短単位区切りと長単位区切りでそれぞれ分かち書きしたテキストを 1 つのファイルに記述し, 同一空間上で学習を行い, 取得した. このとき分散表現の次元数は 50 とした.

表 2 タスク別の訓練事例の件数

データ名	件数
分散表現合成用データ	150,718 件
意味的関係付きデータ	1,673 件
品詞パターン付きデータ	89,951 件

4.2 比較実験

本論文では, 複合語の分散表現を合成する際に, 構成語の品詞パターンの分類タスクをサブタスクとしたマルチタスク学習のモデルを提案し, 河野ら [6] と Komiya ら [3] のモデルと比較する. 以下に, 実験を行った手法を整理する.

手法 1: 基本モデル
図 1 に示したように, 構成語ふたつの分散表現を入力として, 複合語ひとつの分散表現を合成するモデルを作成した. これを基本モデルと呼ぶ.

手法 2: 意味的関係の分類を用いた SVM 分類モデル
Komiya ら [3] は, 構成語間の意味的関係の分類を SVM で行い, そののちにそれぞれの分類ごとに構成語から複合語の分散表現を合成している. 図 3 に分類モデルのイメージを示す.

手法 3: 品詞パターンの分類を用いた SVM 分類モデル

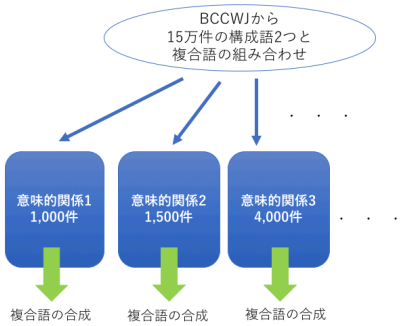

図 3 各意味的関係へ分類後に複合語合成

手法 2 の手順で実験を行う際, 構成語間の意味的関係の分類のかわりに, 品詞パターンの分類を SVM で行った.

手法 4: 意味的関係の分類を用いた MLP 分類モデル
手法 2 の手順で実験を行う際, 構成語間の意味的関係の分類を SVM のかわりに多層パーセプトロンを用いて行った.

手法 5: 品詞パターンの分類を用いた MLP 分類モデル
手法 2 の手順で実験を行う際, 構成語間の意味的関係の分類のかわりに, 品詞パターンの分類を行い, また, SVM のかわりに多層パーセプトロンを用いた.

手法 6: 意味的関係の分類を用いたマルチタスク学習モデル
河野ら [6] は, 図 2 に示したネットワークを使い, 意味的関係の分類をサブタスクとして用いて, 構成語から複合語の分散表現を合成をマルチタスク学習で行った. その後, その学習で得られた重みを使って FT を行い, 構成語から複合語の分散表現を合成を再学習した.

手法 7: 品詞パターンの分類を用いたマルチタスク学習モデル
提案手法である. 手法 6 と同様の手順を使って学習を行うが, マルチタスク学習のサブタスクには, 構成語の意味的関係の分類の代わりに品詞パターンの分類タスクを用いた.

なお, 手法 6 と手法 7 において FT を行う際は, 分散表現合成用データには意味的関係や品詞パターンが付与されていないため, 分類タスクの loss を強制的に 0 とした.

5. 結果

図 4 の構成図の通り, 合成した複合語と教師値である複合語の分散表現を, 二分割交差検証を使い, コサイン類似度で評価した.

提案手法の手法 1〜手法 7 までの結果を表 3 に示す.

手法 6, 手法 7 において FT を行う前のマルチタスク学習における結果をそれぞれ表 4, 5 に示す.

Vol.2020-NL-246 No.23
Vol.2020-SLP-134 No.23
2020/12/3

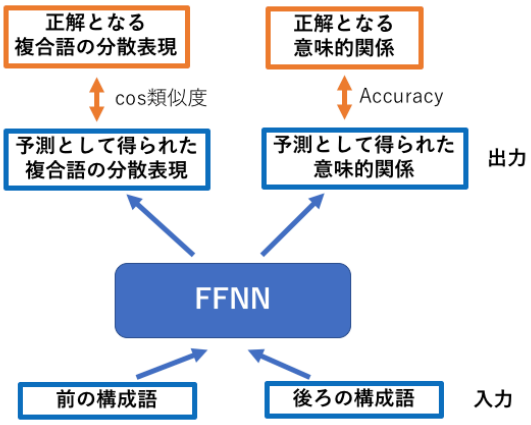

図 4　構成語から複合語と意味的関係を出力するモデル

表 3　手法別の合成性能

手法番号	教師値とのコサイン類似度
1	0.6302
2	0.6467
3	0.6492
4	0.6501
5	0.6434
6	0.6580
7	**0.6618**

表 4　意味的関係付きデータを利用したマルチタスク学習結果

複合語の合成 cos 類似度	意味的関係の分類精度
0.5762	0.8459

表 5　品詞パターン付きデータを利用したマルチタスク学習結果

複合語の合成 cos 類似度	品詞パターンの分類精度
0.6027	0.9207

手法 2〜5 では分類された各意味的関係ごとに複合語の合成を行っている．学習データに意味的関係付きデータを使用したもの (表 6) と，品詞パターン付きデータを使用したもの (表 7) をそれぞれ以下に示す．

6. 考察

表 3 より、手法 1（基本モデル）のコサイン類似度が一番小さいことから，構成語間の意味的関係や，品詞パターンの情報が，複合語の分散表現の合成に有効であることが分かる．また，最もコサイン類似度が高かった手法が手法 7（品詞パターンの分類を用いたマルチタスク学習モデル）であり，次に高かった手法が手法 6（意味的関係の分類を用いたマルチタスク学習モデル）であることから，分類してから別々に学習するよりも，マルチタスク学習を行う方が，複合語の分散表現の合成の精度が高いことが読み取れる．

また，マルチタスク学習に意味的関係付きデータを使った手法 6 に比べ，品詞パターン付きデータを使った手法 7 の方が合成性能が高い結果となった．これは品詞パターン付きデータが，

- 意味的関係付きデータより件数が多い．
- 意味的関係を BCCWJ から採用しているためより正確である．

ことが考えられる．

しかし，各品詞パターン別に MLP と SVM を使って分散表現合成用データを分類したところ (図 7)，MLP・SVM ともに，分類された**品詞パターン 12(名詞-固有名詞-人名+名詞-固有名詞-人名)** の合成性能が低い結果となった．この原因として人物名はパターンが多く，FFNN だけでは推測しにくいことが考えられる．分類後の**品詞パターン 12** において合成性能を上げるための案として，人物の所属ジャンル (スポーツや作家など) や年代のベクトルを付与させることで合成性能が上がるのではないかと考えている．

表 6　意味的関係付きデータで学習後の各意味的関係でのコサイン類似度
() 内は分類件数

意味的関係	マルチタスク学習（手法 4）	SVM（手法 2）
1	0.5041 (30,106)	0.5243 (27,806)
2	0.5005 (738)	0.4768 (2,287)
3	0.5508 (246)	0.5237 (5,502)
4	0.5373 (322)	0.5080 (198)
5	0.6435 (8,724)	0.6471 (9,622)
6	0.5959 (18,484)	0.5223 (8,911)
7	0.5790 (1,121)	0.4918 (2,984)
8	0.7880 (49,762)	0.7704 (56,990)
9	0.4948 (4,850)	0.5389 (3,054)
10	0.6432 (24,024)	0.6283 (32,706)
11	0.6671 (414)	0.5806 (649)
12	0.7051 (6,240)	0.0000 (0)
13	0.5491 (5,687)	0.5117 (9)
マイクロ平均	0.6501	0.6467

表 7　品詞パターン付きデータで学習後の各品詞パターンでのコサイン類似度
() 内は分類件数

品詞パターン	マルチタスク学習（手法 5）	SVM（手法 3）
1	0.5031 (10,381)	0.5038 (12,969)
2	0.6419 (23,798)	0.6747 (38,682)
3	0.5437 (8,194)	0.5817 (14,468)
4	0.4859 (4,392)	0.4938 (4,398)
5	0.4957 (4,081)	0.5015 (4,782)
6	0.6430 (6,234)	0.6544 (9,776)
7	0.5838 (2,457)	0.5781 (2,980)
8	0.5632 (838)	0.5462 (1,027)
9	0.5447 (4,543)	0.5677 (5,814)
10	0.5247 (2,490)	0.5790 (4,563)
11	0.4774 (1,390)	0.5330 (504)
12	0.4147 (713)	0.4678 (862)
13	0.7046 (81,207)	0.7403 (49,893)
マイクロ平均	0.6434	0.6492

7. おわりに

以上の実験結果より, 構成語 2 つからなる複合語の合成において, BCCWJ から取得した品詞情報を意味的関係として分類するサブタスクを用いることで, 合成性能が上がることが確認できた.

久本ら [7] は単語の合成に双方向 LSTM を使用している. 本研究は FFNN だけであったため, 今後双方向 LSTM などの構成のニューラルネットワークを使用することで合成性能向上が考えられる.

また, 本研究では構成語 2 つから複合語を合成しているが, 構成語 3 つ以上から複合語を合成することについても, BCCWJ の品詞情報を利用して合成を行うことが今後の課題である.

参考文献

[1] Kazuma Hashimoto and Yoshimasa Tsuruoka. Adaptive joint learning of compositional and non-compositional phrase embeddings. In *Proceedings of the 54th ACL*, pp. 205–215, 2016.

[2] Teruo Hirabayashi, Kanako Komiya, Masayuki Asahara, and Hiroyuki Shinnou. Composing word vectors for japanese compound words using bilingual word embeddings. *PACLIC 2020*, 2020. no. 21.

[3] Kanako Komiya, Takumi Seitou, Minoru Sasaki, and Hiroyuki Shinnou. Composing word vectors for japanese compound words using dependency relations. *CICLING 2019*, 2019. no. 229.

[4] Masayasu Muraoka, Sonse Shimaoka, Kazeto Yamamoto, Yotaro Watanabe, Naoaki Okazaki, and Kentaro Inui. Finding The Best Model Among Representative Compositional Models. In *Proceedings of PACLIC 2014*, pp. 65–74, 2014.

[5] Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. Mimicking word embeddings using subword rnns. *EMNLP*, pp. 102–112, 2017.

[6] 河野慎司, 古宮嘉那子. マルチタスク学習を利用した短単位の分散表現から長単位の分散表現の合成. 言語処理学会第 26 回年次大会, 2020.

[7] 久本空海, 山村崇, 勝田哲弘, 竹林佑斗, 高岡一馬, 内田佳孝, 岡照晃, 浅原正幸. chive: 製品利用可能な日本語単語ベクトル資源の実現へ向けて. 第 16 回テキストアナリティクス・シンポジウム, pp. 40–45, 2020.