国立国語研究所学術情報リポジトリ

All-words WSDとfine-tuningを利用した分類語彙表の語義の分散表現の構築

著者	柳沼 大輝，古宮 嘉那子，新納 浩幸
雑誌名	言語資源活用ワークショップ発表論文集
巻	4
ページ	179-184
発行年	2019
URL	http://doi.org/10.15084/00002568

All-words WSD と fine-tuning を利用した 分類語彙表の語義の分散表現の構築

柳沼 大輝 (茨城大学大学院理工学研究科) *
古宮 嘉那子 (茨城大学大学院理工学研究科) †
新納 浩幸 (茨城大学大学院理工学研究科) ‡

Generating Sense Embeddings of word List by Semantic Principles Using All-words WSD and Fine-tuning

Daiki Yaginuma (Ibaraki University)
Komiya Kanako (Ibaraki University)
Hiroyuki Shinnou (Ibaraki University)

要旨

近年、単語を低次元の実数値ベクトルで表した分散表現が自然言語処理の様々な分野で利用されている。さらに、単語の分散表現や、その作成手法の応用により語義ごとの分散表現を作成する研究がされており多くのタスクで有効な結果を残している。しかし、一般に人手で語義が付与されたコーパスは量が少ないため、十分な量の語義が付与されたコーパスの確保は困難である。そこで、本稿では、語義を自動的に付与した大量の精度が低いコーパスから、作成した分散表現をもとに、人手で語義が付与された少量の精度が高いコーパスを用いて fine-tuning を行い、分類語彙表の語義の分散表現を作成し、その精度を検証した。その結果、分散表現の距離を用いた順位付けによる評価により、fine-tuning による精度の向上が認められた。

1. はじめに

本論文では、コーパスの全単語を対象とした語義曖昧性解消 (All-wordsWSD) を用いて自動的に語義を付与したコーパスと、人手で語義が付与されたコーパスの二種類のコーパスを使用し、fine-tuning を用いて分類語彙表(1) の語義の分散表現を作成する手法を提案する。

近年、単語を低次元の実数値ベクトルで表した分散表現 (Word Embeddings) が自然言語処理の様々な分野で注目されている。さらに、単語の分散表現や、作成手法の応用により語義ごとの分散表現 (Multi-sense Embeddings) を作成する研究がされており、多くのタスクで有効な結果を残している。一方、単語の分散表現は、通常、単語列であるテキストコーパスを用いて構築される。語義を付与されたコーパスから語義列を作成し、単語列の代わりに語義列を用

* 18nm740n@vc.ibaraki.ac.jp
† kanako.komiya.nlp@vc.ibaraki.ac.jp
‡ hiroyuki.shinnou.0828@vc.ibaraki.ac.jp
(1) https://pj.ninjal.ac.jp/corpus_center/goihyo.html

いれば、単語の分散表現と同様の手法で語義の分散表現の構築が可能である。しかし、一般に人手で語義が付与されたコーパスは量が少ないため、十分な量の語義列コーパスの確保は困難である。また、All-wordsWSD の手法を用いて語義の付与を行う研究 (鈴木類ほか (2018)) は行われているため、その結果を用いれば語義列コーパスは入手可能であるものの、常に正しい語義が付与されるとは限らず、分散表現の作成には適さない可能性がある。

本研究では、All-wordsWSD の手法を用いて語義を付与した大量の精度が低いコーパスから作成した分散表現をもとに、人手で語義が付与された少量の精度が高いコーパスを用いて fine-tuning を行い、分類語彙表の語義の分散表現を作成した。また、fine-tuning の効果を調べるため、それぞれのコーパスから作成した分散表現及び、元となる分散表現と、コーパスの組み合わせを入れ替えた分散表現を作成した。得られた分散表現を分類番号が木構造であることを利用し、分散表現同士の距離を用いて順位付けによる評価を行った。

実験の結果、双方の分散表現の作成における fine-tuning の有用性が確認された。

2. 関連研究

近年、単語を低次元のベクトルで表した表現である、分散表現が自然言語処理の様々な分野で幅広く利用されている。さらに、新たな単語の分散表現や、単語の分散表現の作成手法を応用した語義ごとの分散表現作成の研究がされており、多くのタスクで有効な結果を残している。大内らの研究 (大内ほか (2018)) では、それぞれの語義の類義語の分散表現を、語義の分散表現として利用している。山木らの研究 (山木翔馬ほか (2017)) では、MSSG モデルを用い、教師データを利用して語義の頻度情報を考慮した語義の分散表現の構築手法を提案している。ただし、これらの研究では通常の単語のコーパスや分散表現を用いるのみで、語義が付与されたコーパスを使用していない。通常、単語の分散表現は、単語列であるテキストコーパスを用いて構築される。語義を付与されたコーパスから語義列コーパスを作成し、単語列の代わりに語義列を利用することで単語の分散表現と同様の手法を用いて語義の分散表現の構築が可能である。しかし、一般に語義の付与は人手で行われるため、十分な量のコーパスを確保するのは困難である。

鈴木らの研究 (鈴木類ほか (2018)) では、All-wordsWSD を各単語に語義を与えるラベリング問題とみなし、語義を付与したコーパスの作成を行った。このように全単語に自動に語義を付与したコーパスを利用して語義列を作成し、語義列から語義の分散表現を作成することも可能である。しかし、自動に付与された語義は、人の手を用いて付与されたものとは違い、常に正しい語義が付与されるとは限らず、分散表現の作成には適さない可能性がある。

以上の観点から、本研究では、自動タグ付けの手法によって語義を付与した大量のコーパスをもとに分散表現を作成し、人の手によって語義が付与された少量のコーパスを用いて fine-tuning を行うことで、精度の高い分類語彙表の語義の分散表現を作成を目指す。

3. 提案手法

本研究では、分散表現の作成にあたり、現代日本語書き言葉均衡コーパス (BCCWJ) をもとに二種類の語義を付与して作成した二つのコーパスを利用した。一つ目は、All-wordsWSD

の手法を用いて自動で語義を付与した大量の語義列コーパス (以下、All-wordsWSD コーパス) であり、作成には、先行研究 (鈴木類ほか (2018)) により構築された All-wordsWSD システムを利用した。二つ目は人手で語義を付与した少量の語義列コーパス (以下、人手コーパス) であり、国立国語研究所による分類語彙表番号アノテーションデータを利用した (加藤祥ほか (2017))。単語列コーパスと語義列コーパスの具体例を表 1 に示す。

表 1　語義列コーパスの具体例

モノ で なく 心 で は ない の か
1.4000 で 3.1200 1.3000 で は 3.1200 の か

また、それぞれのコーパスの単語数、語彙数、語義数を表 2 に示す

表 2　コーパスの単語数と語彙数と語義数

対象の分散表現	単語数	語彙数	語義数
All-wordsWSD コーパス	23,968,826	75,028	851
人手コーパス	347,094	3,164	916

本研究では、語義の分散表現の作成に word2vec[(2)] を用いた。また、作成済みの分散表現を初期値として与え、新たなコーパスを使って分散表現を訓練しなおすという手順で fine-tuning を行った。

分散表現は以下の四種類を作成した。一つ目は All-wordWSD コーパスから作成した分散表現の「All-wordsWSD ベクトル」である。二つ目は All-wordsWSD ベクトルを初期値として与え、人手コーパスに対して fine-tuning を行い作成した「All-wordsWSD-fine ベクトル」である。また、比較のために人手コーパスから同様の手法で分散表現を作成し、All-wordWSD コーパスに対して fine-tuning を行い分散表現の作成を行った。この時、人手コーパスから分散表現を作成した分散表現を「人手ベクトル」、人手ベクトルを初期値に All-wordsWSD コーパスで fine-tuning を行った分散表現を「人手-fine ベクトル」とする。

4. 評価実験

分散表現の評価のため分類語彙表を用いて作成した分散表現の妥当性の検証を行った。 分類語彙表は、語義が階層構造（木構造）の中で定義されている概念辞書であるため、同じノードに属する語義同士は距離が近くなることが予想される。これを利用することで、作成した分散表現の評価を行うことが可能である。評価は以下の手順で行う。

(1) 作成した単語 w_i の分散表現 e_i 毎に、分類語彙表中の対応するノード n を求める

(2)n の、兄弟ノード集合 N を得る

(3)n 中の w_i 以外の全単語の分散表現と e_i の距離を測りその平均を得る

(2) https://code.google.com/archive/p/word2vec/

(4)N 中のノード毎に、ノード中の全単語の分散表現と e_i の距離を測りその平均を得る
(5)e_i がどのノードと最も近いかの順位を得る
(6) 最も e_i と近いと判定されたノードの、正解のノードの距離の差を求める

4.1 実験設定

word2vec の計算に使用されるパラメータには、次元数を 200、ウィンドウサイズを 5、バッチサイズを 1000、反復回数を 5 とし、アルゴリズムには cbow を利用した。また、fine-tuning の際、使用した word2vec の訓練用パラメータは事前に分散表現を作成したときと同じ条件を用いた。作成した分散表現の距離の比較にはコサイン類似度を利用した。

4.2 実験結果

表 3 に本手法で作成したそれぞれの分散表現の属するノードの順位及び、

表 3　距離を用いた順位付けによる評価

対象の分散表現	順位の平均	一位との差の平均	葉の数
All-wordsWSD ベクトル	6.868	0.102	42
All-wordsWSD-fine ベクトル	3.143	0.049	42
人手ベクトル	2.945	0.059	42
人手-fine ベクトル	2.644	0.046	42

最も近い葉とされたノードと正解のノードの距離の差について全分散表現の平均を示す。表 3 から、それぞれの分散表現の「順位の平均」について、最も低い順位が All-words WSD ベクトルの 6.868 であることが分かる。葉の数は平均で 42 あるので、ランダムに所属ノードを推測した場合には 21 位程度と予想されるため、本研究において作成した分散表現に妥当性があることが分かる。

5. 考察

本研究において作成した分散表現は、人手ベクトル、All-wordsWSD ベクトル共に「順位の平均」及び「一位との差の平均」の評価値が、fine-tuning によって向上している。両手法ともに、分散表現の fine-tuning による精度向上が確認できた。

また、人手ベクトルによる分散表現の評価値は、All-wordsWSD ベクトルによる分散表現の評価値を上回っている。同様に、人手ベクトル-fine の評価値は、All-wordsWSD-fine ベクトル評価値を上回っている。これは、fine-tuning を行わない分散表現を作成するコーパスとしては人手コーパスの方が優れていることを示唆しており、良質なコーパスは fine-tuning において、初期値の学習に使用することでより精度の高い分散表現が作成できると推測される。

一方で、本手法で提案した All-wordsWSD-fine ベクトルの評価値は、従来のコーパスを使用している人手ベクトルの評価値を下回る結果となった。しかし、本手法において分散表現作成の際、コーパスの比較のため、word2vec のパラメータは全て一律で行っている。コーパス

のサイズや特徴により、細かな調整を行うことで分散表現の精度が向上する場合がある。例えば、反復回数を 10 回にして作成した All-wordsWSD ベクトルと人手ベクトルの評価結果は表 4 のようになる。表 4 の結果では、表 3 の結果を下回っている。それぞれのコーパスに適した

表 4　反復回数を 10 回にして作成した分散表現の評価結果

対象の分散表現	順位の平均	一位との差の平均	葉の数
All-wordsWSD ベクトル	7.52	0.105	42
人手ベクトル	3.217	0.043	42

パラメータ設定については広範な実験が必要である。しかし、All-wordsWSD コーパスを人手ベクトルの fine-tuning に利用する手法が最もよかったことから、人手コーパスのパラメータ調整として有効であることが分かった。

本研究での実験結果では、fine-tuning に関わらず人手コーパスを用いて作成した分散表現の評価値が、All-wordsWSD コーパスを用いて作成した分散表現の評価値を上回っていたことから、分散表現の作成については人手コーパスの方が優れているという結果になった。ただし、分散表現の作成の際、対象コーパスの大きさが、対象単語の語義の精度よりも、作成される分散表現の精度に必ずしも関わるとは言い切れない。例えば、All-wordsWSD コーパスは、すべての単語に対する正解率は 8 割程度、多義語に対する正解率は 7 割程度、単語数については約 6,7 倍程度であったことが挙げられる (表 2)。どの程度の大きさのどの程度の精度の自動で語義を付与したコーパスが、どの程度の大きさの人手で語義を付与したコーパスに匹敵するか今後も調査を進めていきたい。

6. おわりに

本研究では、All-wordsWSD コーパスをもとに作成した分散表現を人手コーパスを用いて fine-tuning を行い、分類語彙表の語義の分散表現の作成を行った。結果として、双方のコーパスをもとにした場合でも分散表現の作成における fine-tuning の精度向上が確認された。また、少量の人手によるコーパスを All-WordsWSD によってタグ付けした大量のコーパスによって fine-tuning する手法が、本実験中で最も高い精度となった。今後は、分散表現を作成する際に必要なコーパスの語義の付与の精度や大きさの重要度の比較や、パラメータを調整した実験等、分散表現の精度の向上などを課題とし、検証を行っていきたい。

文　献

鈴木類・新納浩幸・古宮嘉那子 (2018).「双方向 LSTM による分類語彙表番号を語義とした all-words WSD」　国語研言語資源活用ワークショップ, pp. 2–4.

大内克之・新納浩幸・古宮嘉那子・佐々木稔 (2018).「類義語を利用した単語の分散表現から語義の分散表現の構築」　言語処理学会第 22 回年次大会, pp. 99–102.

山木翔馬・新納浩幸・古宮嘉那子・佐々木稔 (2017).「教師データを用いた語義の分散表現の構築」　言語処理学会第 23 回年次大会発表論文集, pp. 78–81.

加藤祥・浅原正幸・山崎誠 (2017).「現代日本語書き言葉均衡コーパス」 に対する分類語彙表番号アノテーション. 言語処理学会第 23 回年次大会発表論文集, pp. 306–309.