E-025

日本語 WordNet における語義・概念の分散表現獲得
Extracting Embedded Vectors for Lexemes and Synsets of Japanese WordNet

國府大輝 † Daiki Ko　　**竹内孔一** † Koichi Takeuchi

1. はじめに

辞書資源は，意味役割付与 [1] 等の自然言語処理分野のタスクに利用されている．これらのタスクの性能は辞書資源の持つ語彙に大きく制限される．すなわち，タスクの性能向上のためには，既存の辞書資源を更新・拡張していく必要がある．しかし人手で辞書の語彙を増やすことは，多大なコストと時間を要するため，辞書資源の更新を自動化する研究が行われてきた．また辞書資源の自動更新には分散表現を利用するものが多く存在する [2]．その分散表現獲得手法として，近年 word2vec[3] や fasttext[4] 等が注目されており，文書分類 [5] などの自然言語処理のタスクに適用されている．しかし多くの分散表現獲得手法は，多義語の場合にも 1 つの単語に 1 つの分散表現を割り当てるため，各語義に対応した分散表現を得る事は出来ない．また，複数語義の分散表現は分散表現を獲得するために用いたコーパスに強く影響されるという特徴がある．この問題を解決することが出来れば，語義曖昧性解消 [6] 等の意味タスクにおいて有用であると考えられる．

そこで本稿では，辞書資源を利用し日本語の単語・語義・概念の分散表現を同一な空間上に構築する．

2. 語義・概念の分散表現

本研究では，辞書構造を反映した分散表現を獲得する手法として提案されている AutoExtend[7] を利用する．AutoExtend は階層的な語義関係を構築している WordNet に適用して分散表現を獲得する．そこで，同様の構造を持つ，日本語 WordNet[8] を利用して，日本語の辞書における概念体系を考慮した分散表現の獲得を目指す．また本研究で AutoExtend に与える日本語 WordNet の概念間の関係は，上位・下位関係と近似関係である．

2.1. AutoExtend

AutoExtend は既存の単語の分散表現を入力とし，辞書資源における辞書構造を反映した自己符号化器により，当該の辞書資源における語義・概念に対して分散表現を与える．単語・語義・概念の分散表現を同一な空間上に構築するため，異なる種類の分散表現を直接比較する事が可能となる．

より具体的には，まず訓練済みの単語の分散表現を入力とし，その分散表現を単語の持つ語義の分散表現に分解する．次に語義の分散表現を用いて，語義の集合である概念に分散表現を与える．最後に，これらの変換の逆の操作を行い，圧縮した分散表現を元の単語の分散表現に復元する．このようにして中間層にあたる語義の分散表現を調整して獲得する．全体としては，語義を介して単語と概念間の変換を行うモデルとなっている．

2.1.1. AutoExtend を構成する数式

AutoExtend を数式を用いて説明する．AutoExtend が用いる辞書資源の情報構造では，単語はいくつかの語義を持ち，これらの語義の集合として概念が構成される．そこで AutoExtend の Encoder 側における単語 $w^{(i)}$ の概念 $s^{(j)}$ に属する語義を $l^{(i,j)}$ としたとき，辞書資源の情報構造は式 (1)(2) のように表せる．

$$w^{(i)} = \sum_j l^{(i,j)} \tag{1}$$

$$s^{(j)} = \sum_i l^{(i,j)} \tag{2}$$

単語から語義に変換する対角行列を $E^{(i,j)}$ とする．

$$l^{(i,j)} = E^{(i,j)} w^{(i)} \tag{3}$$

式 (2)(3) より

$$s^{(j)} = \sum_i E^{(i,j)} w^{(i)} \tag{4}$$

となる．また，式 (4) を単語と概念全体に拡張するとシンプルなテンソル積（クロネッカー積）の形で書ける．

$$S = \boldsymbol{E} \otimes W \tag{5}$$

式 (5) において，S は概念の分散表現，W は単語の分散表現をそれぞれ並べた行列であり，$\boldsymbol{E}$ は $E^{(i,j)}$ を並べた 4 階テンソルである．式 (5) が AutoExtend の Encoder にあたる．また，AutoExtend の Decoder 側における概念 $s^{(j)}$ の単語 $\overline{w}^{(i)}$ に属する語義を $\bar{l}^{(i,j)}$ としたとき，辞書資源の情報構造は式 (6)(7) のように表せる．

$$s^{(j)} = \sum_i \bar{l}^{(i,j)} \tag{6}$$

$$\overline{w}^{(i)} = \sum_j \bar{l}^{(i,j)} \tag{7}$$

概念から語義に変換する対角行列を $D^{(j,i)}$ とする．

$$\bar{l}^{(i,j)} = D^{(j,i)} s^{(j)} \tag{8}$$

式 (7)(8) より

$$\overline{w}^{(i)} = \sum_j D^{(j,i)} s^{(j)} \tag{9}$$

†岡山大学大学院自然科学研究科 Graduate School of Natural Science and Technology Okayama University

となる．また，式 (9) も式 (4) と同様に単語と概念全体に拡張するとシンプルなテンソル積の形で書ける．

$$\overline{W} \quad = \quad \boldsymbol{D} \otimes S \tag{10}$$

式 (10) において，$\overline{W}$ は単語の分散表現をそれぞれ並べた行列であり，$\boldsymbol{D}$ は $D^{(j,i)}$ を並べた 4 階テンソルである．式 (10) が AutoExtend の Decoder にあたる．

$$\underset{\boldsymbol{E},\boldsymbol{D}}{\operatorname{argmin}} \|\boldsymbol{D} \otimes \boldsymbol{E} \otimes W - W\| \tag{11}$$

全体としては，式 (11) のように W を復元するモデルとなっている．

3. 語義・概念の分散表現の評価

獲得した分散表現に対して評価実験を行う．獲得した分散表現にはテキストベースの単語の分散表現と意味的に正しく対応する語義・概念が存在するかどうかにより学習の良さを評価する．本稿では，単語「美味しい」に対して分散表現間類似度の高い語義・概念を上位 10 件示す．そして，概念の場合は日本語 WordNet 中のその概念の定義文を示し，語義の場合はその語義に紐付けられている概念の定義文を示す．

3.1. 評価結果

単語「美味しい」に対しての結果を表 1 に示す．また，表 2 には単語「美味しい」と類似度の高い順に上から表 1 中の概念に対する定義文を示す．

これらの表から AutoExtend は辞書資源に対して適切な分散表現を獲得出来ていると言える．表 1 で最も多く出力されている概念ラベルである 01586342-a は，対応する表 2 の定義文を見ると単語「美味しい」に紐付けられるべき概念である事がわかる．「食べて美味しい」や「見て美味しい」と言った日本語特有の「美味しい」に対する考え方が反映された概念であると言えるからである．

表 1: 「美味しい」と類似する語義・概念

順位	語義・概念
1	美味しい-01586342-a
2	おいしい-02395115-a
3	美味い-01808671-a
4	旨い-00604617-a
5	美味い-01123148-a
6	旨い-02226162-a
7	02395115-a
8	01586342-a
9	好いたらしい-01586342-a
10	絶品-03727605-n

表 2: 日本語 WordNet における概念の定義文

概念ラベル	概念の定義文
01586342-a	性質または外観が快活な、愉快な、あるいは感じのよい
02395115-a	味覚が良い
01808671-a	感覚に気持ちよい
00604617-a	快適さまたは目的またはニーズに適している
01123148-a	特に指定したものに適している、望ましい、あるいはプラス方向の品質であるさま
02226162-a	知識、技能、および才能を持っているか、または示すさま
03727605-n	芸術家や工芸家の最も優れた作品

3.2. 議論

3.1 節の結果から AutoExtend による語義・概念の分散表現構築は有用であることがわかった．そのため，この概念の分散表現を使い辞書資源の新語登録に利用できると考えられる．例えば未知語に対しコーパスを用いて分散表現を与え，その未知語と辞書資源におけるすべての概念の分散表現との類似度を算出し，未知語との類似度が上位である概念を未知語と紐付けることで辞書資源の新語登録が実現できる．この際，紐付ける概念を上位何語にするか，類似度の高い概念がなかった場合はどうするか等の問題が今後の課題となる．

4. おわりに

本稿では，AutoExtend に既存の単語の分散表現と辞書資源である日本語 WordNet を与え語義・概念の分散表現を構築し，単語「美味しい」に対して意味的に近い語義・概念が出力されることを確認した．このことにより，AutoExtend による日本語の語義・概念の分散表現構築の有用性について確認することが出来た．今後はこの概念の分散表現を利用して辞書資源の新語登録も行う予定である．

謝辞

本研究の遂行にあたり JSPS 科研費 JP19K00552 の助成を受けた．

参考文献

[1] 岡村拓哉, 竹内孔一, 石原靖弘. ニューラルネットワークを利用した日本語意味役割ラベル付与システムの構築. 言語処理学会第 24 回年次大会発表論文集, pp. 101–104, 2018.

[2] 金田健太郎, 小林哲則, 林良彦. 語義・概念の分散表現を利用した Semantic Taxonomy Enrichment. 言語処理学会第 24 回年次大会発表論文集, pp. 793–796, 2018.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, 2013.

[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. FastText.zip: Compressing Text Classification Models. *arXiv preprint arXiv:1612.03651*, 2016.

[5] 難波英嗣. 人工知能による文書分類. 情報の科学と技術, Vol. 66, No. 6, pp. 277–281, 2016.

[6] Roberto Navigli. Word Sense Disambiguation: A Survey. Vol. 41. ACM, 2009.

[7] Sascha Rothe and Hinrich Schütze. AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. In *Proceedings of the Association for Computational Linguistics*, 2015.

[8] 栗林孝行, Francis Bond, 黒田航, 内元清貴, 井佐原均, 神崎享子, 鳥澤健太郎. 日本語ワードネット 1.0. 言語処理学会第 16 回年次大会発表論文集, pp. 978–981, 2010.