K-027

語彙学習のための Word2Vec を用いた単語固有の意味抽出手法の提案と検証

Proposal and Validation of a Word-Specific Semantic Extraction Method Using Word2Vec for Vocabulary Learning

角田　悠翔†　　前田　大輝†　　小尻　智子‡
Yuto Kakuta　　Daiki Maeda　　Tomoko Kojiri

1. はじめに

類義語を適切に使い分けるためには，個々の単語の意味だけではなく，それらの単語の相違を正確に理解する必要がある．例えば，「案外」と「意外」は両方とも「予想・期待と実際が違っていること」を表しているが，「案外」は「予想外ではあるが，強く驚いていないこと」であり，「意外」は「強い驚きや予想とのギャップを表すこと」を表している．そのため，驚きの強さに応じて使うべき語を変える必要がある．本研究の目的は，このように共通の意味が存在する 2 つの語に対して，それぞれの固有の意味単語を提示することで適切な語の使用を支援する．

単語を学ぶ際によく用いられる辞書には個々の単語の意味は詳細に記載してあるが，単語間の相違には焦点をあてていない．類義語辞典からは意味が類似している語を把握することはできるが，それらの違いはわからない．人が使用方法を迷う単語の組は様々であるため，あらかじめすべての単語の固有の意味を用意しておくことは現実的ではない．そのため，指定された単語の組のそれぞれに対して，他方の単語にはない固有の意味単語を自動的に抽出し，学習者に提示することができれば，適切な語の使用を支援できる．

これまで様々な語彙の学習支援システムが開発されている[1]．これらは個々の単語の意味の理解を支援対象としており，2 つの語の相違には焦点があたっていない．学習者が作成した文章に含まれる文脈にそぐわない単語を検出し，それに対して類似した意味を持つより適切な代替語を複数提示するシステムも開発されている[2]が，類義語間の意味の相違を明示的に学習者に示す支援はしていない．

一方，Word2Vec などの機械学習モデルを用いることで，単語は多次元のベクトルとして表現できる．それぞれの単語は意味的な類似性に応じて互いに相対的な位置関係を持つように分布している．本研究では，このベクトル空間における類義語の分布に着目し，類義語同士の語彙ベクトルの中間領域には，それらの共通の意味を持つ語が分布しやすく，一方で各語に固有の意味を表す語は，他方の語から距離の離れた方向に分布する傾向があると仮定する．この仮定に基づき，Word2Vec を用いて他の語とは異なる領域の語を抽出し，類義語の固有の意味を表す単語として提示する方法を提案する．

2. Word2Vec

Word2Vec とは，単語の意味的特徴を多次元の連続空間上のベクトルとして表現するためのモデルであり，意味の類似した単語ほど空間上で互いに近接するように学習される[3]．語彙ベクトルのコサイン類似度をとることで，語彙同士の類似を表現できる．また，語彙ベクトルを四則演することも可能である．例えば「王様」－「男性」+「女性」の結果は，「女王」を表す語彙ベクトルと類似したベクトルとなる．

† 関西大学大学院理工学研究科　Graduate School of Science and Engineering, Kansai University
‡ 関西大学システム理工学部　Faculty of Engineering Science, Kansai University

3. Word2Vec を用いた固有の意味単語の抽出手法

固有の意味単語は，その単語に単語固有の成分を加えたもので表現できると考えられる．その単語固有の成分は，その単語から類義語との共通の成分をひいたものとして考えられる．よって，式(1)，(2)から成る，固有の意味単語の抽出手法を提案する．図 1 に式(1)，(2)が表すベクトルの概念図を示す．

$\vec{w_i}$は単語 i の語彙ベクトルを表す．$\vec{f_i}$は単語 i 固有の成分を表すベクトルであり，$\vec{E_i}$は単語 i の固有の意味単語である固有強調ベクトルを示す．式(1)は固有強調ベクトルを求める式であり，固有の意味単語を求めたい単語のベクトルに，その単語固有の成分を一定数強調したものをたしあわせたものとなっている．その単語固有の成分は，式(2)のように，その単語のベクトルから類義語との平均ベクトルをひいたもので求められる．

$$\vec{E_i} = \vec{w_i} + k\vec{f_i} \quad (k\text{は任意定数}) \tag{1}$$

$$\vec{f_i} = \left\{\vec{w_i} - \frac{(\vec{w_i} + \vec{w_j})}{2}\right\} \tag{2}$$

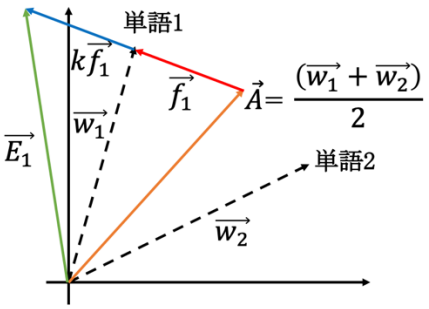

図 1　式(1)，(2)のイメージ図

4. 検証実験

提案手法の有効性を評価するための実験を行った．本実験の Word2Vec の学習には Wikipedia コーパスを用いた．学

習には gensim ライブラリを用い，ベクトルの次元数を 300，エポック数を 9 に設定した．

4.1 平均ベクトルの検証実験 1

平均ベクトル$(\overrightarrow{w_1}+\overrightarrow{w_2})/2$が 2 つの類義語の共通の意味を表現しているかを検証するために実験を実施した．2 組の単語の共通の意味単語は，それぞれの類義語のうちの共通の類義語で表現され，固有の意味を表す語彙は異なる類義語であるとみなすことができる．平均ベクトルの検証実験では，提案手法が導出した平均ベクトルとコサイン類似度が大きい上位 15 単語（平均ベクトル周辺単語）の中でコサイン類似度が 0.6 以上の単語の中に，共通の類義語が 1 つ以上含まれているかによって評価した．

検証に用いた単語の組は，コトバンクのデジタル大辞泉[4]で共通の意味単語が存在する類義語 45 組である．このうち，平均ベクトル周辺単語の中でコサイン類似度が 0.6 以上の単語に共通の類義語が 1 つ以上あったものは 26 組あった．表 1 に用いた単語群を示す．共通の類義語が存在していたものには，ID の下に下線をひいている．［結末・結果］のように，同漢字が 1 つ以上含まれている類義語では，25 組中 20 組で共通の意味単語が抽出できていた．したがって，提案手法は同漢字を含む場合に有効である可能性があると考えた．

表 1　実験で用いた類義語 45 組

ID	単語組	ID	単語組	ID	単語組	ID	単語組
1	感動・感心	13	収集・蓄積	25	経歴・履歴	37	採用・採択
2	推定・推測	14	連絡・接続	26	迅速・敏捷	38	完了・達成
3	出版・刊行	15	器具・道具	27	信頼・信用	39	真実・事実
4	購入・購買	16	辞書・辞典	28	探索・調査	40	役割・責任
5	解答・回答	17	危険・危惧	29	堅実・地道	41	定義・意味
6	準備・用意	18	心配・不安	30	記入・記載	42	主義・理念
7	形態・形状	19	感触・感覚	31	落胆・失望	43	知能・頭脳
8	早急・緊急	20	重要・大事	32	実行・実施	44	注意・配慮
9	承認・承諾	21	企画・計画	33	残念・遺憾	45	結果・結末
10	意図・目的	22	綿密・厳密	34	経費・費用		
11	提示・表示	23	適切・適当	35	案外・理解		
12	避難・逃避	24	業務・作業	36	禁止・規則		

4.1.2 平均ベクトルの検証実験 2

同漢字が 1 文字以上含まれる類義語 31 組を対象に，4.1.1 と同様の検証実験を実施した．結果を表 2 に示す．平均ベクトル周辺単語とそれらのコサイン類似度が 0.6 以上の単語の中に共通の類義語が存在する組は 28 組であり，90.3%であった．これらのことより，同漢字が 1 つ以上含まれる類義語の組では，本提案の想定どおり，平均ベクトルが共通の意味単語を含むと言える．

表 2　同漢字が 1 文字以上含まれる類義語 31 組

ID	単語組	ID	単語組	ID	単語組	ID	単語組
1	発刊・刊行	9	使用・利用	17	輸送・運送	25	維持・保持
2	損失・損害	10	模写・描写	18	問題・疑問	26	目的・目標
3	法律・法令	11	団結・結束	19	発展・伸展	27	描画・描出
4	類似・酷似	12	停止・中止	20	調査・精査	28	回復・復元
5	推定・推測	13	調子・様子	21	活用・利用	29	戦闘・戦争
6	順序・順番	14	理論・論理	22	要件・条件	30	意識・認識
7	勉強・勉学	15	隠蔽・隠匿	23	管理・監理	31	悲惨・惨劇
8	欲求・欲望	16	感動・感慨	24	合致・一致		

4.2 固有強調ベクトルの検証実験

提案する固有強調ベクトルが，類義語の固有の意味単語を表現するかを検証するための評価実験を実施した．検証に用いた類義語は，同漢字が 1 つ以上含まれる類義語 43 組である．そのうち，2 つの単語両方に固有の意味単語が存在した組（両単語組）が 18 組，片方だけに固有の意味単語が存在した組（片単語組）が 25 組であった．定数kの値を d 1/2，1，5，10，50 と変化させ，そのときの固有強調ベクトルを表す単語とコサイン類似度が大きい上位 15 個（固有強調ベクトル周辺単語）が，導出したい固有の意味単語を含むかどうかを検証した．

表 3 に，固有強調ベクトル周辺単語が固有の意味単語を含んでいた単語の割合を，k の値ごとに示す．両単語組は，1 つの組に単語が 2 つあるとみなすことができるため，18 組の類義語で合計 36 個の単語のうち，該当する単語を含む割合を示す．表 3 より，片単語組よりも両単語組の方が固有の意味単語を導出できた割合が多いことがわかる．さらに，kの値が 1/2 の時に最も割合が高くなっていた．固有の意味単語が導出できた両単語組の例として，類義語［決断・決定］では，［決断］の固有の意味としては［決意］が，［決定］の固有の意味として［確定］が導出できた．
今後はどのような類義語で固有の意味単語が導出できるかを今後分析していく必要がある．

表 3　固有強調ベクトルの周辺単語が固有の意味単語を含んでいる単語の割合

k	1/2	1	5	10	50
片単語組	0.20	0.20	0.16	0.08	0.05
両単語組	0.64	0.42	0.39	0.36	0.28

5. おわりに

本研究では，Word2Vec を用いて表現された語彙ベクトルを用いて，与えられた類義語の組の固有の意味単語を導出する手法を提案した．検証実験の結果，同漢字が 1 つ以上含まれる類義語では平均ベクトルが共通の意味単語を表現できることが明らかになった．しかし，固有の意味単語を導出できたのは，最大で 64%しかなかった．

今回の検証では両単語組と比較すると片単語組の精度が非常に悪かった．この理由について，今後より多くの類義語を用いて実験をすることで，明らかにしていきたい．

さらに，今回の検証結果をふまえ，提案手法を用いて適切な語の使用方法の理解をどのように支援していくのか検討したい．

参考文献

[1] A. Toniolo, K. C. Pucihar, M. Kljun, “VocabulARy: Learning Vocabulary in AR Supported by Keyword Visualisations”, IEEE Transactions on Visualization and Computer Graphics, Vol.28, No.11, pp.3748-3758 (2022).
[2] C. Wang, S. Mao, T. Ge, W. Wu, X. Wang, Y. Xia, J. Tien, D. Zhao, “Smart Word Suggestions for Writing Assistance”, arXiv preprint arXiv:2305.09975v1, pp. 1-12 (2023).
[3] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient Estimation of Word Representations in Vector Space”, arXiv preprint arXiv:1301.3781, Vol. 3, pp. 1-12 (2013).
[4] コトバンク：デジタル大辞泉キーワード一覧，https://kotobank.jp/dictionary/daijisen/.