対訳文を用いた同義語・類義語・対義語の抽出

柳原弘哉 [1] 村上仁一 [2]
[1] 鳥取大学大学院 持続性社会創生科学研究科 工学専攻
[2] 鳥取大学工学部
[1]m22j4052m@edu.tottori-u.ac.jp
[2]murakami@tottori-u.ac.jp

概要

従来，同義語・類義語・対義語は，意味に基づいて人の手によって分類される [1]．そのため，収集におけるコストが高い．また，自動的なシノニム抽出の研究 [2][3] は数多く存在する．しかし，同義語・類義語・対義語が区別されず，総括してシノニムと扱われる．そこで本研究では，コーパス内の同義語・類義語・対義語を区別した形で自動抽出することを目的とする．通常，シノニムは意味に基づいて収集されるが，コーパスに「意味」という情報は存在しない．そこで，「翻訳の対応関係」を「意味」と仮定し，対訳コーパスからシノニム抽出を行う．

実験の結果，同義語 94.3%，類義語 78.2%，対義語 60.7% の精度が得られた．また，提案手法より対義語において辞書に未記載のシノニム抽出にも成功した．

1 はじめに

同義語・類義語・対義語は単語間の関係性を表現し文意理解において重要である．対義語も一部類似した性質を持つことから便宜上，本論文において同義語・類義語・対義語を総合してシノニムと表現する．従来，シノニムは意味に基づいた人手による判断で分類 [1] された．シノニム単語を収集したシノニム辞書が制作に長い月日を必要とすることから人手分類のコストの高さを示している．

一方で，シノニムを自動抽出する研究 [2][3] は数多くされており手法も様々である．しかし，同義語・類義語・対義語にそれぞれ区別して抽出している研究はあまり見られない．同義語・類義語と対義語は置き換えることで文意が正反対になるため，区別が特に重要であると考えられる．そのため，本論文では，共通の対訳コーパスからシノニムを区別した形で自動的に抽出をすることを試みる．

2 従来手法

2.1 人手による収集

人手収集におけるシノニムは意味を判断基準としており，各シノニムの意味は辞書 [1] によって次のように定義されている．

同義語：語形は異なるが意義はほぼ同じ言葉
類義語：意味の類似する単語
対義語：意味の上で互いに反対の関係にある語

2.2 自動抽出の研究

2.2.1 パターンを用いたシノニム抽出

Chklovski ら [2] は，類似する単語が特定の文法パターンで共起する性質に注目した．WordNet から関連性の高い動詞ペアを収集し，事前に定義されたパターンを使用することで動詞におけるシノニム抽出を行った．

2.2.2 分散表現を用いたシノニム抽出

Li ら [3] は，類似する単語が単語分散表現の意味空間において近接する性質に注目した．Word2Vec[4] を使用することで単語間の類似性を計算し，スペクトルクラスタリングによって単語をクラスタリングすることでシノニム抽出を行った．

3 問題点・目的

自動的なシノニム抽出の研究は，人手によるコストが不要という利点がある．しかし，人手収集とは異なりシノニム内の明確な区別がないことが多い．その理由として，本来なら意味基準で分類するシノニムを自動抽出においては意味以外の情報を利用しており，シノニム間の境界の曖昧性を解消できないためだと考えられる．曖昧性の一つとして，対義語対はカテゴリーが一部共通する性質から類似性を持つことが挙げられる．例えば，“白” と “黒” は，色の明暗から対義語として扱われるが，色という同じカテゴリーであるため類似する単語とも解釈できる．

以上のことを踏まえて，本研究ではシノニムの自動抽出において同義語・類義語・対義語を区別した形で抽出することを目的とする．

4 提案手法

言葉は「意味」によって単語自体が持つ概念や性質といった知識を他人と共有することができる．知識の共有ができる観点から「翻訳の単語対応」は「意味」と同等と考えることができる．例えば，“服”という単語は「身につけるもの．きもの．」といった意味であるが，日本語を知らない英語話者に対しては対訳単語である “clothes” を伝えることで，“服”という単語が持つ概念を共有することができる．この性質から，共通する翻訳を持つ単語同士は意味が同じと仮定することで，類似する単語の抽出に対訳単語を利用できると考えられる．

一方，類似単語の一部とみなせる対義語は，同一文において置き換えることで正反対の内容を表現できる．例えば，“右” と “左” の対義語対において「交差点を右に曲がる．」は “右” を “左” に置き換えることで正反対の文になる．つまり，文脈 (周囲の単語) によって類似性を求められる分布仮説において，対義語が最も類似すると考えられる．

4.1 再定義

本研究では，「翻訳における単語対応」を「意味」とみなすため，2 言語の翻訳対応を利用したシノニムの再定義を行う．2 言語は日本語-英語である．

4.1.1 同義語

- 単一カテゴリー内に存在する単語対
- 日本語単語の対が英語訳において完全に共通
  例）“病気” と “病”
  病気＝ disease ＝病，病気＝ illness ＝病

4.1.2 類義語

- 単一カテゴリー内に存在する単語対
- 日本語単語の対が英語訳において一部共通
  例）“青” と “緑”
  青 =green ＝緑，青＝ blue ≠緑

4.1.3 対義語

- 単一のカテゴリー内に存在する単語対
- 日本語単語の対が英語訳において共通しない
- 共通する文脈で置き換え可能
  例）“右” と “左”
  右＝ right ≠左，左＝ left ≠右
  右に曲がる． ， 左に曲がる．

4.2 変換テーブル

4.1 節の 2 言語を利用した再定義を踏まえると，シノニムの関係に有る単語対それぞれには対訳単語が存在する．つまり，シノニム抽出には日本語単語 2 対，英語単語語 2 対の計 4 対が必要である．そこで，変換テーブル [5] を利用する．

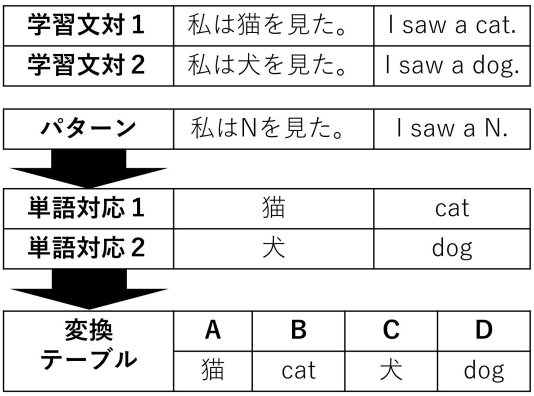

図 1 変換テーブル

変換テーブル [5] は単語 4 組の関係性を定義するテーブルである．パターンが共通する 2 組の対訳文において，パターンの変数に対応する単語が抽出されることで生成される．各単語の相対性により「A が B ならば C は D」という関係が成立する．図 1 の例では，「猫が cat ならば犬は dog」である．

4.3 Word2Vec

Word2Vec[4] は文脈 (周囲の単語) を用いて単語を数値化することでベクトル表現を可能にするニューラルネットワークの手法である．対義語同士が類似する文脈で使用される性質から，Word2Vec[4] において対義語同士の類似度は最も高いと考えられる．

5 実験方法

5.1 実験手順

5.1.1 同義語

単語対が翻訳において完全に共通する組み合わせを抽出する．(表 3 参照)

1. 変換テーブル ABCD で英語 B，D が同一単語，且つ日本語 A，C が別単語の組み合わせを選択
2. 翻訳が複数存在することを考慮して，1. で得られた日本語単語 A について変換テーブルを利用し，対応する英単語をすべて収集⇒集合 A’
3. 2. と同様の処理を 1. で得られた日本語単語 C についても行う⇒集合 C’
4. 2. と 3. で得られた集合 A’ と集合 C’ を比較し，完全に一致する組み合わせを同義語として出力

5.1.2 類義語

単語対が翻訳において一部共通する組み合わせを抽出する. (表 4 参照)

1. 変換テーブル ABCD で英語 B，D が同一単語，且つ日本語 A，C が別単語の組み合わせを選択
2. 翻訳が複数存在することを考慮して，1. で得られた日本語単語 A について変換テーブルを利用し，対応する英単語をすべて収集⇒集合 A'
3. 2. と同様の処理を 1. で得られた日本語単語 C についても行う⇒集合 C'
4. 2. と 3. で得られた集合 A' と集合 C' を比較し，共通しない部分がある組み合わせを類義語として出力

5.1.3 対義語

日本語・英語同士が最も類似する組み合わせを Word2Vec を利用して抽出する．(表 5 参照)

1. 変換テーブル ABCD で英単語 B，D が別単語であり，日本語単語 A，C も別単語である組み合わせを選択 (A と B は対訳単語である)
2. Word2Vec を利用し，1. で得られた日本語単語 A に最も類似する単語を抽出⇒ A'
3. Word2Vec を利用し，1. で得られた英語単語 B に最も類似する単語を抽出⇒ B'
4. 変換テーブル中から ABA'B' となる組み合わせを検索
5. 1.~4. の処理を CDC'D' の組み合わせでも行う

5.2 実験データ・条件

表 1 使用データ

変換テーブル	701,828
日本語単文データ	163,188
英語単文データ	163,188

5.2.1 変換テーブル

森本 [5] が作成した変換テーブルを使用する．森本は，電子辞書などの例文から抽出された 163,188 文の単文対訳コーパス [6] を利用して変換テーブルを作成した．生成された変換テーブルは 701,828 組である．

5.2.2 Word2Vec

Python の gensim[7] を使用する．データの統一を取るために，変換テーブルと同じ 163,188 文の単文対訳コーパス [6] で学習を行う．各パラメータは，window=5，size=10,000 とし，英語モデル，日本語モデルをそれぞれ作成した．

6 実験結果

同義語・類義語・対義語の各抽出数と人手評価による精度を表 2 に示す．ただし，漢字や片仮名を含む表記ゆれや数字などのノイズは評価に考慮しない．評価者は著者 1 人である．

表 2 抽出結果

	人手評価	正解率
同義語	50 / 53	94.3%
類義語	1,097 / 1,403	78.2%
対義語	54 / 89	60.7%

出力例の一部を表 3~5 に示し，変換テーブルが参照した対訳文を例文として示す．

6.1 同義語

表 3 同義語

A	B	C	D	評価
価格	price, prices the prices,	値段	price, prices the prices	○
間違い	error, mistakes errors, mistake	ミス	error, mistakes errors, mistake	○
秩序	order	順序	order	×

例文)

1).価格と値段
野菜の価格が急騰している.
The price of vegetables is soaring.
野菜の値段が下がる.
The price of vegetables drops.

2).秩序と順序
それ は 新た な 秩序 の 到来 を 告げ た .
That ushered in a new order .
逆 の 順序 に おく . Place in the reversed order .

6.2 類義語

表 4 類義語

A	B	C	D	評価
お昼	lunch, noon	正午	at noon, midday, noon	○
和平	peace, The peace	平和	Peace, peace peaceful	○
板	board, plank platform	委員会	board, meeting committee	×

例文)

1).お昼と正午
お昼 まで に は , まだ 少し 間 が ある .
There's still a little time left until noon .
正午 に 汽笛 が 鳴る . The whistle blows at noon .

2).板と委員会
板 が 歪む． The board is distorted .
委員 会 は 明日 開か れる 予定 だ .
The board is meeting tomorrow .

6.3 対義語

表 5 対義語

A	B	C	D	評価
東	east	西	west	○
無罪	innocence	有罪	guilt	○
雪	snow	雨	rain	×

例文）

1).東と西
風は東へ吹いている．The wind is blowing east .
風が西へ吹く． The wind blows west .
2).雪と雨
雪 が やん だ . The snow has stopped .
雨 が やん だ . The rain has stopped .

7 考察

7.1 不正解の調査

実験結果における不正解対の原因調査により，方式の限界による要因と変換テーブル作成時のパターンの誤りによる要因に大別された．同義語は方式限界のみであった．各要因の割合を表 6 に示す．

表 6 不正解の割合

	方式限界	(割合)	誤パターン	(割合)
類義語	35 / 84	(41.7%)	49 / 84	(58.3%)
対義語	23 / 35	(65.7%)	12 / 35	(34.3%)

方式限界による要因

同義語・類義語では，英単語は共通するが類似しない日本語対が抽出された．この問題の原因は単語の持つ多義性である．言語間の概念の差異や文脈により単語の用途が変化するため，方式限界であると考えられる．(表 3，表 4 の不正解例を参照)

対義語では，反対の意味が存在しない日本語対が抽出された．この問題の原因は単語に反対が存在するかの区別が困難なことである．変換テーブル自体はパターンが共通する組み合わせを出力しており，組み合わせ自体に類似性は存在しないため，方式限界であると考えられる．(表 5 の不正解を参照)

パターンの誤りによる要因

森本 [5] の変換テーブルはパターンを利用して作成されており，パターンの誤りが単語対応の誤りに影響を与えている．つまり，パターンの精度を上げることで類義語・対義語については抽出精度が向上すると考えられる．

7.2 辞書による評価 (対義語)

実験結果より同義語・類義語と比較し，対義語の正解率が低い．原因として，7.1 節で述べた反対の存在しない単語を抽出していることが考えられる．そこで，対義語辞典において対義語と定義される単語対に限定して再度精度を調査した．対義語辞書には，weblio 対義語反対語辞典 [8] を利用した．

表 7 明確に対義語である単語の調査

	辞書評価	(正解率)	人手評価	(正解率)
対義語	30 / 45	(66.7%)	34 / 45	(75.6%)

調査の結果，対義語辞書 [8] において不正解だが，人手評価において正解と判断された対が得られた．また，調査の過程で，実験結果における表 7 に含まれない対義語が得られた．つまり，対義語辞書において定義されないが，人手評価において正解と判断される対義語である．それぞれの一部を表 8，表 9 に示す．

表 8 辞書で不正解となった対義語

A	B	C	D	辞書
先生	teacher	生徒	pupils	先生 ⇔ 弟子
朝	morning	夜	evening	朝 ⇔ 夕，晩
円	yen	ドル	dollars	円 ⇔ 方

表 9 辞書で定義されない対義語

A	B	C	D
犬	dog	猫	cat
女王	Queen	国王	King
東京	Tokyo	大阪	Osaka

この結果から，本研究の提案手法は辞書に記載されていない対義語を抽出できたことを示している．

8 まとめと今後の課題

従来の研究では，意味基準によって区別されるシノニムの曖昧性が解消されない問題があった．そこで，本研究では，「翻訳の単語対応」を「意味」と仮定することでシノニムを自動抽出・分類する方法を提案した．実験の結果，同義語 94.3%，類義語 78.2%，対義語 60.7% の精度が得られた．また，対義語において辞書に未記載の単語抽出といった有効性を得られた．

今回，提案した手法では抽出数が不十分であると考えられる．そのため，精度を損なわずに抽出数を増加させることを今後の課題とする．

参考文献

[1] 新村出, 新村出記念財団. 広辞苑. 岩波書店, 第 6 版, 2008.

[2] Timothy Chklovski and Patrick Pantel. VerbOcean: Mining the web for fine-grained semantic verb relations. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 33–40, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[3] Li Zhang, Jun Li, and Chao Wang. Automatic synonym extraction using word2vec and spectral clustering. In **2017 36th Chinese Control Conference (CCC)**, pp. 5629–5632, 2017.

[4] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In **International Conference on Learning Representations**, 2013.

[5] 森本世人. 類似度を利用した変換テーブルの精度向上. 言語処理学会 第 27 回年次大会, 2021.

[6] 村上仁一, 藤波進. 日本語と英語の対訳文対の収集と著作権の考察. 第一回コーパス日本語学ワークショップ, 2012.

[7] Radim Rehurek and Petr Sojka. Gensim–python framework for vector space modelling. **NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic**, Vol. 3, No. 2, 2011.

[8] GRAS グループ株式会社. weblio 対義語反対語辞典, (2023-01 閲覧). https://thesaurus.weblio.jp/antonym/.