Word2Vec と対訳単語対を利用した対義語の自動抽出

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

対義語抽出は自然言語処理分野において非常に重要なタスクである．しかし，対義語抽出に関する研究 [1][2] は少なく，WordNet 等の人手で作成された辞書を正解ラベルとして利用する手法が主流である．本研究では，コーパスにおける “文脈情報” と “単語の対訳関係” を利用することで，正解ラベルを使用せず全自動（人手作成の辞書に依存しない方法）で対義語抽出を試みた．コーパスの文脈情報には，日本語 Wikipedia の記事で学習した Word2Vec[3][4] モデルを使用した．また，単語の対訳関係には，JParaCrawl に対して IBM モデルに基づく FastAlign[5] で単語対応付けを行った．テスト実験により，59% の正解率が得られた．

1 はじめに

本研究は，対義語抽出に焦点を当てる．しかし，類義語についても言及する．

類義語と対義語の抽出・分類は自然言語処理分野において重要なタスクである．類似性・対称性等の言葉の関係性を原理的に解明することで言語の理解や処理技術の向上に貢献する．しかし，類義語抽出に関する研究が多い反面，対義語抽出に関する研究は少ない．その理由として，対称の関係にある単語を抽出するタスクが非常に困難であることが挙げられる．そのため，対義語抽出に関する他の研究では，正解ラベルを必要とする教師あり学習の手法が主流であり，人手で作成した類義語・対義語の辞書が利用される．

本研究では，対義語の性質に基づいた手法を利用することで正解ラベルに依存しない全自動の対義語抽出を提案する．抽出の手法として，「**類義語・対義語の文脈の類似性**」と「**意味と翻訳の対照性**」の 2 点の性質に着目する．

2 関連研究

2.1 人手作成のパターンによる関係語抽出

Chklovski ら [1] は，人手で取得された特定の関係を示す文法パターンを利用して，類義語と対義語を含む 5 種類の異なる関係語の抽出を試みた．文法パターンの利用により語義リソースに依存せず，web コーパスを活用して広範囲な動詞の関係語を抽出した．

2.2 教師データによる類・対 義語の分類

Samenko ら [2] は，類義語と対義語の教師データで学習した単語埋め込みモデルにおいて，類似する単語の距離が小さく，対称の単語の距離が大きくなるよう最適化することで，類義語と対義語の分類可能性を調査した．また，提案手法により現代の単語埋め込みモデルには類義語と対義語を区別できる情報が含まれていることを示した．

3 問題点

対義語に関する研究は，類義語に関する研究と比較して少ない．その理由として，対称の関係にある単語を抽出するタスクが非常に困難だからと考えられる．そのため，対義語抽出に関する他の研究では，正解ラベル付きの学習データを使用した教師あり学習の手法が主流である．しかし，正解ラベルは人手に依存しており作成する上でコストが問題となる．また，教師モデルは単に正解ラベルのパターンを学習する傾向にあるため，言語の原理や構造を理解することが困難な可能性がある．

4 目的

本研究では，正解ラベルを使用しない手法を提案することで，対義語抽出のタスクに対する原理的なアプローチを試みる．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

5 類義語・対義語の性質

5.1 類・対 義語の文脈の類似性

類義語対・対義語対同士は，単語の意味カテゴリが共通するため類似性を持つ．加えて，類似する単語同士は置き換えられる可能性があり，類義語対・対義語対は前後の文脈が類似する性質を持つ．

類義語対は，共通する性質を示す単語対であり，単語を構成する意味的な要素が共通する．例えば，“喜び” と “幸せ” はポジティブな感情や心の状態を表しており，共通する意味カテゴリである．しかし，完全には共通せず，一部異なるニュアンスを含む．例えば，“喜び” と “幸せ” では，“喜び” は，外部からの刺激で引き起こされる比較的短い期間の感情を表現する傾向がある．対して，“幸せ” は，持続的で定常的な感情を表現する傾向にある．つまり，意味カテゴリの共通性が単語の類似性を示す．また，分布仮説より，カテゴリが共通する (意味が類似する) 単語は文脈が類似する性質を持つ．一方，対義語対は，対称の性質を示す単語対である．しかし，単語の要素全てが対称の関係ではない．むしろ，大部分の要素は類似しており，意味カテゴリも類義語対と同様に共通する．例えば，対義語である “白” と “黒” は明暗の対比で対称関係にあるが，色・光度という共通する意味カテゴリに属する．つまり，対義語対は意味カテゴリが共通する (意味が類似する) ため，文脈が類似する性質を持つ．

5.2 意味と翻訳の対照性

「意味」と「翻訳」には対照性がある．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」と「意味」には対応関係がある．例えば，“本” という単語は「書籍．書物．[1)]」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 “book” を伝えることで，“本” という単語が持つ概念を共有することができる．また，翻訳自体が言語を横断して同じ内容を伝達する性質上，“書籍：book” と “本：book” の様に翻訳が共通する単語同士は意味が同じである．

6 提案手法

6.1 再定義

本論文では，5 章で説明した性質に基づき類義語・対義語を表 1 に定義する．

1) 広辞苑 第 6 版，岩波書店，2008 年

表 1 類義語と対義語の定義

類義語	対義語
• 意味カテゴリ 共通 • 文脈が類似 • 翻訳 (=意味) が**同じ**	• 意味カテゴリ 共通 • 文脈が類似 • 翻訳 (=意味) が**違う**
例) “**病気**” と “**病**” **病気**にかかる． **病**にかかる． **病気** = disease = **病**	例) “**左**” と “**右**” **左**に曲がる． **右**に曲がる **左** = left ≠ right = **右**

6.2 抽出手法

6.1 の再定義に基づき，**文脈情報**と**対訳単語対**を利用した対義語抽出の手法を提案する．文脈情報には Word2Vec[3][4]，対訳単語対の取得には FastAlign[5] をそれぞれ利用する．

6.2.1 文脈情報

対義語対は文脈が類似する．そこで，周囲の単語 (文脈) を考慮する単語埋め込みモデルを利用する．

Word2Vec[3][4] は，ターゲット単語の周囲に出現する単語を統計的に処理して，単語のベクトル表現を可能にする．文脈が類似する単語対はベクトル空間内における距離が近いため，cos 類似度が高い．

6.2.2 対訳単語対

類義語・対義語の分類に対訳関係を利用する．対訳単語の取得において，完璧な精度を実現することは不可能である．しかし，6.2 の再定義では，対義語は対訳単語が共通しないと定義づけており，翻訳精度が結果に大きく影響する．そこで，誤りを前提として対訳の**共通割合**を考える．大規模な対訳データにおいて，日本語単語対の各対訳に占める共通する対訳単語の割合で対義語を識別する．

FastAlign[5] は，IBM model 2 のパラメータを簡素化して計算効率を向上させた統計的機械翻訳の手法であり，単語アライメントの取得に特化している．

6.3 抽出手順

手順を以下に示す．J1 と J2 は日本語単語対であり，cos 類似度が互いに最大となる単語対を選択する．また，E1 は J1，E2 は J2 の対訳英単語である．

1. テストデータにおいて cos 類似度が相互に最大となる J1，J2 を抽出
2. J1，J2 の対訳単語 E1，E2 を FastAlign から取得
3. E1，E2 が共通しない J1，J2 を**対義語**として出力

この方法により，文脈が類似 (=意味カテゴリ共通) し，翻訳 (=意味) の異なる単語対の抽出が可能となる．抽出の例を図 1 に示す．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

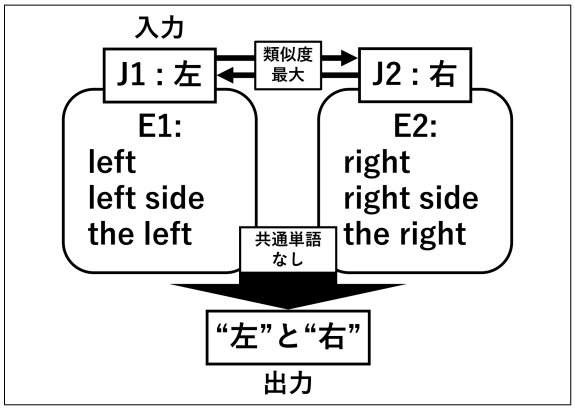

図 1 “左” と “右” の抽出例

7 実験設定

7.1 実験条件

データベースにおける日本語文は全て mecab-python3 1.0.6 で単語に分割した．

Word2Vec パラメータは，vector_size=200，window_size=10 とした．学習データには，2023/9/11 における日本語 Wikipedia の記事を使用した．

FastAlign パラメータは，デフォルトを使用した．学習データには，bicleaner の値が 0.70 以上の JParaCrawl[6] の対訳コーパスを使用した．

テストデータ 日本語文の分割を行い，単語を取得した．データには，電子辞書等から抽出して作成された日英単文対訳文 [7] の日本語文を使用した．

フィルタリング・枝刈り

- Word2Vec の類似度に閾値を設定
- 対訳単語対の共通割合に閾値を設定
- 対義語になり得ない英数字の除外
- 平仮名のみの単語を除外し，異表記対の削減
- 頻度が少なく，信頼性の低い対訳単語の除外

以上の設定は DEV データからの類推により人手で決定する．また，共通割合については，J1 と J2 の両方が閾値を満たす AND と，J1 と J2 の一方が閾値を満たせば良い OR の 2 種類の方式で実験した．

7.2 使用データ

使用するデータベースを半分に分割し，各データは DEV 実験とテスト実験で使用する．DEV 実験は，テスト実験における閾値を決定するため行う．使用したデータベースを表 2，データベースを使用して作成した対訳単語・テストデータを表 3 に示す．また，対訳単語対は JParaCrawl から生成し，テストデータは日英単文対訳文から生成した．(7.1 参照)

表 2 データベース

日本語 Wiki	全データ	1,385,000記事
	DEV 実験	693,590記事
	テスト実験	693,590記事
JParaCrawl	bicleaner0.7 以上	18,383,212 文
	DEV 実験	9,191,606 文
	テスト実験	9,191,606 文
日英単文対訳		161,338 文

表 3 処理済みデータ

対訳単語対	DEV 実験	206,458,091 対
	テスト実験	206,464,657 対
テストデータ		43,148単語

8 実験結果

8.1 テスト実験

DEV 実験において正解率の最も高かった共通割合 40%以下，cos 類似度 0.95 以上で実験を行い，結果を人手で評価した．評価者は著者 1 名である．実験条件と実験結果を表 4 に，出力例を表 5 に示す．

表 4 実験条件と実験結果

共通割合	cos 類似度	方式	出力数	正解率
40%以下	0.95 以上	OR	**44**	**59%**

表 5 出力例

J1	J2	評価	類似度	J1 共通	J2 共通
右舷	左舷	○	0.965	14.3%	26.3%
偶数	奇数	○	0.953	24.8%	81.4%
東岸	西岸	○	0.953	29.0%	7.7%
先輩	後輩	○	0.968	65.5%	31.8%
母方	父方	○	0.975	7.4%	6.5%
貸方	借方	○	0.955	0%	0%
火曜	木曜	×	0.978	0%	0%
少佐	中佐	×	0.955	12.8%	27.7%
筋骨	隆々	×	0.966	14.5%	100%
脚注	使い方	×	0.992	0.9%	0.02%

8.2 不正解対の考察

テスト実験における不正解の原因を以下に示す．

連続する概念 曜日，階級，漢数字等の連続性を持つ概念の単語対が出力された．概念が連続する単語は，単語自体が持つ意味よりも前後の概念の相対的な関係性が重要視されると考えられる．つまり，単語を置き換えても文全体の内容に影響せず，同じ文脈で使用できる可能性が高い．そのため，cos 類似度が高くなる．且つ，各単語は独立した概念で翻訳が共通しないことが理由と考えられる．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

連続語　“筋骨” と “隆々”，“脚注” と “使い方” の様に複合名詞，または助詞を挟んで連続する名詞の対が出力された．連続語同士は，文内における位置が近い．そのため，文脈で共起する単語が類似し，cos 類似度が高くなることが理由と考えられる．特に，“脚注” と “使い方” は，Wikipedia のフレーズ「脚注の使い方」で頻出するため，Word2Vec のモデルに使用したコーパスも影響すると推測される．

また，統計的機械翻訳の手法を利用しているため，本来なら翻訳が共通しない単語対であっても，共起する性質により連続語の対訳単語が共通することも理由と考えられる．特に，“隆々” は JParaCrawl において 2 文の例外を除く全ての文で，“筋肉” または “筋骨” と共起していたため，対訳単語が高確率で “筋骨” と共通すると推測される．

9　議論

実験結果より，“**文脈情報**” と “**単語の対訳関係**” を利用することで，正解ラベルを必要とせず，対称な関係にある単語対を自動的に抽出できることが示された．しかし，精度・出力数という観点から現在の手法には改善する余地があると考える．本章では，DEV 実験の結果を踏まえて提案手法の性質について言及し，今後の課題に触れる．

9.1　DEV 実験

対訳単語の共通割合を 30%，40%以下の 2 種類，Word2vec の cos 類似度を 0.85，0.90，0.95 以上の 3 種類で実験を行った．評価者は著者 1 名である．表 6 における曖昧については 9.1.1 で説明する．

表 6　DEV 実験結果

共通割合	cos 類似度	方式	出力	人手評価　(100 対)	
				正解率	曖昧
30% 以下	0.85 以上	AND	421	33 %	12
		OR	676	40 %	9
	0.90 以上	AND	119	43%	13
		OR	188	41 %	8
	0.95 以上	AND	23	**65%**(15/23)	0
		OR	29	62%(18/29)	0
40% 以下	0.85 以上	AND	510	30 %	9
		OR	816	34 %	3
	0.90 以上	AND	151	49 %	13
		OR	223	45 %	11
	0.95 以上	AND	26	**69%**(18/26)	0
		OR	36	**69%**(25/36)	0

9.1.1　対義語ではないが対称性を持つ単語対

評価が難しい単語対を “曖昧な単語対” と判断した．人手評価では正解に含めないが，解釈次第では対称性を持つと考えることができ，提案手法が対称の単語を抽出できる性質を示す．以下は例である．

協奏曲とソナタ：ピアノ・ヴァイオリンがメインのクラシック音楽．協奏曲がオーケストラ形式に対し，ソナタはソロ形式であるため，相補的である．

ヤンキースとドジャース：MLB の名門チーム．リーグは異なるが，ライバル関係にある．拠点が東岸と西岸であり，共に金満球団という観点で対比されるため，対称的である．

9.1.2　提案手法の性質

表 6 より，提案手法では cos 類似度の値が結果に著しく影響しており，出力数と精度の間にはトレードオフの関係が確認できる．さらに，DEV 実験では，人名や地名を含む対義語になり得ない固有名詞が出力されており，今後の実験で不正解として影響する可能性が考えらる．

9.2　今後の課題

対義語ではない**連続する概念・連続語・固有名詞対**を除外するフィルターを設定することで，誤り率の軽減が考えられる．さらに，**閾値・パラメータのチューニング**により精度向上が期待できる．DEV 実験を 6 種類の閾値で実施したが，網羅的に検証することで，より適切な設定が得られる可能性がある．加えて，今回の実験では，Word2Vec や FastAlign のパラメータは単一の設定で実験したため，チューニングすることでモデルの改善も考えられる．

本研究では，対義語の抽出可能性と精度に主眼を置いた．そのため，正解のサンプル数が少なく，考察が不十分な可能性がある．また，対義語のカバー率についても考慮しておらず，今後の課題である．

10　まとめ

本論文では，“文脈情報” と “対訳単語対” を利用した対義語抽出の手法を提案した．「**類義語・対義語の文脈の類似性**」と「**意味と翻訳の対照性**」の 2 点の性質に着目することで，正解ラベルを必要としない全自動の抽出を試みた．テスト実験により 44 対の対義語が得られ，59%の正解率であった．しかし，使用データの最適化，実験のパラメータ・条件の調整，新たなフィルターを導入等の改善により，今後さらなる精度向上が期待できる．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] 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.

[2] Igor Samenko, Alexey Tikhonov, and Ivan P. Yamshchikov. **Intuitive Contrasting Map for Antonym Embeddings**. IOS Press, October 2021.

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

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

[5] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[6] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In **Proceedings of The 12th Language Resources and Evaluation Conference**, pp. 3603–3609, Marseille, France, May 2020. European Language Resources Association.

[7] 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.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).