2021年度（令和3年度）　卒業論文

変換テーブルを用いた
同義語･類義語･対義語の抽出

指導教員

村上仁一
村田真樹

鳥取大学工学部　電気情報系学科
自然言語処理研究室
B18T2119U 栁原 弘哉

概要

同義語 · 類義語 · 対義語は, 単語同士の特定の関係性を表現し文意を理解する上で大いに役立つ. 従来, 同義語 · 類義語 · 対義語は人の手作業によって分類するため, 長い月日をかけて収集されてきた. そのため, 辞書作成の段階では時間的なコストを必要とした. そこで, コンピュータを用いて同義語 · 類義語 · 対義語を機械的に抽出することで問題解決を試みる.

本研究では, 分類する方法として 2 言語間の翻訳対応関係について注目する. また, 機械的に分類する上で関係性を明確に区別するために同義語 · 類義語 · 対義語を再定義する. 以上の方法を踏まえて, パターンにより対訳単語の生成を可能とする変換テーブル [1] を使用して辞書作成を行う. 調査結果より, 機械的な分類が可能であることが分かった. また,[6][7] に存在しない同義語 · 類義語 · 対義語の単語対を抽出することに成功した.

目 次

図 目 次

表 目 次

第1章　はじめに

同義語 · 類義語 · 対義語は単語同士の特定の関係性を表現し, 文意を理解する上で大いに役立つ. 従来, 言葉とは概念や性質といった「意味」によって考えられてきた. 同様に, 同義語 · 類義語 · 対義語の関係も意味に基づいて考えられるため, 人手による分類が不可欠となる. しかし, 人の力では膨大な量の語彙を一朝一夕に収集, 分類することはできない. そのため, 辞書作成においては時間的コスト・経済的コストが必要とされた. そこで, 本研究では変換テーブルを使用して機械的に同義語 · 類義語 · 対義語を分類することで辞書作成を自動的に行いコストの削減を行う.

本論文の構成は以下の通りである. 第 2 章では, 従来定義について述べる. 第 3 章では, 提案手法について述べる. 第 4 章では, 再定義について述べる. 第 5 章では, 実験概要について述べる. 第 6 章では, 実験結果について述べる. 第 7 章では, 考察について述べる. 第 8 章では, 今後の課題について述べる. 第 9 章では, 本研究のまとめについて述べる.

第2章　従来手法

2.1　辞書による定義

2.1.1　同義語

- 広辞苑 [2]

  語形は異なるが意義はほぼ同じ言葉.「即刻」と「即時」の類. 同意語. シノニム.

- 明鏡国語辞典 [3]

  語形は異なるが中心的な意味が同じである語.「みち」と「道路」,「やまい」と「病気」の類. 同意語. シノニム⇔対義語

- ブリタニカ国際大百科事典 [4]

  ⇒類義語

2.1.2 類義語

- 広辞苑 [2]

  意義の類似する単語.「おこる」と「いかる」,「両親」と「父母」など. 類語

- 明鏡国語辞典 [3]

  意味がよく似ている二つ以上の語.「時間」と「時刻」,「話す」と「語る」,「こわい」と「おそろしい」など. 類語

- ブリタニカ国際大百科事典 [4]

  一言語体系の⇒語彙のなかで, 互いによく似た意味をもつ 2 つ以上の単語. 同義語. 同意語とも呼ばれるが,「ねじる」と「ひねる」のように用いられる文脈が違うので(「スイッチをねじる」は不可）実際には意味が違うといわざるをえないものや,「食べる」と「食う」のように文体差があるものや,「タレント」と「芸人」のように情意的語感の異なるものなどが多く, 完全に入替のきく同意語というものはまれである, 類義語後の成立には⇒外来語が大きな役割を果たし, 日本語では漢語と和語, 英語ではアングロ · サクソン系の単語とフランス語系の単語で対をなしているものが多い（「旅」と「旅行」,freedom と liberty など）.

2.1.3 対義語

- 広辞苑 [2]

  意味の上での互いに反対の関係にある語.「上」と「下」,「積極」と「消極」の類. 対語（たいご · ついご）. 反対語. 反意語. 反義語. アントニム. ⇔同義語

- 明鏡国語辞典 [3]

  同一言語の中で, その意味が対の関係, あるいは正反対の関係にある語.「右-左」「出席-欠席」「明るい-暗い」「売る-買う」などの類. アントニム. 反意語. 反義語. 反対語. 対語. ⇔同義語

- ブリタニカ国際大百科事典 [4]

  対義語, 反対語ともいう. 互いに反対の意味を持つ一対の語.「行く-来る」「長い-短い」,”bad-good”などがその例であるが,「泣く-笑う」などのように, 必ずしも正反対ではないものも含まれる. したがって語によっては反義語がいくつもある場合があり, 反義語のない場合もある. また, ある語の否定が必ずしもその語の反意語となるとはかぎらない. また, 反対の意味といっても, 意味論的には種々の場合が含まれる.

2.2 問題点

従来辞書において, 言葉とは単一言語 (日本語) で概念や性質といった「意味」によって定義されてきた. それにより, 同義語 · 類義語 · 対義語といった関係性も意味 · 性質に基づいて考えられてきた. しかし, 意味基準の分類では同義語 · 類義語 · 対義語の収集に人手が不可欠であり, 膨大な量の語彙を人手で分類するには多大な時間的 · 経済的コストが必要になる.

第3章　提案手法

本章では, 提案手法について記述する.

3.1　概要

翻訳とは, 文章を同等の内容を表す別言語に置き換える行為であり,2 言語の翻訳における対応する文は同じ意味である. これは単語においても言えることであり,2 言語間で対応する単語同士は同等の内容を表す. ただ, 単語の場合は単語同士の対応が 1 対 1 とは限らない. 一つの単語に対し複数の対応する翻訳が存在する可能性が高い. 翻訳の対応は同等であるため, 複数の対応する翻訳同士は同等の内容を表す別表現と考えられる. よって, 本研究では, 共通する他言語訳 (本研究では英語を使用する.) を持つ日本語単語は同義語もしくは類義語の可能性があると考える. また, 対義語は単語対同士が反対の意味を表しているが, 同じ文脈で使用されることが多い. よって, 類似する文に出現しやすい単語対は対義語の可能性があると考えられる.

第4章　再定義

本章では, 本研究における再定義についての説明を行う. 英語話者に対して日本語の”赤”という情報を伝えるには,”赤”に対応する英単語を使用して説明するのが妥当である. つまり,2 言語間における翻訳の対応関係は, 単一言語における意味と同等であると考えることができる. 従来の同義語 · 類義語 · 対義語は単一言語において意味基準で定義されてきたが, 本研究では 2 言語における翻訳の対応関係に注目して再定義を行う. 使用する言語は日本語と英語である.

4.1　同義語

- 単一のカテゴリー内に存在する単語対
- 日本語単語のペアの英語訳が完全に一致する.

例）病気＝ disease, illness, 病＝ disease, illness
馬鹿＝ fool, idiot, 阿保＝ fool, idiot

4.2　類義語

- 単一のカテゴリー内に存在する単語対
- 日本語単語のペアの英語訳が一部一致する.

例）青=green,blue, 緑＝ green
援助=help,assitance, 救助=help

4.3 対義語

- 単一のカテゴリー内に存在する単語対
- 日本語単語のペアの英語訳が一致することはない.
- 意味が対照的な一対の単語対

例）右＝ right ≠左, 左＝ left ≠右　(方向的に対照)
青＝ green ≠赤, 赤＝ red ≠青　(信号機の意味で対照)

第5章　実験概要

本研究における実験では, 日本語-英語間の対訳単語の関係性を表すデータとして変換テーブル [1] を用いて同義語 · 類義語 · 対義語を分類する. また, 対義語については類似する構文で出現する単語を抽出するために Word2Vec を用いる.

5.1　実験目的

人手による同義語 · 類義語 · 対義語収集では時間的 · 経済的コストがかかる. そこで, 変換テーブルを使用して機械的に収集 · 分類することで時間とコストの削減を試みる. また. コストを削減した方法で自動で辞書作成することを本研究の目的とする.

5.2 変換テーブル

変換テーブルとは, 対訳文同士の相対性によって対応する単語の関係性を定義するテーブルであり,2 つの対訳文に共通するパターンから 2 単語対を抽出する. これは,「A が B ならば C は D である」という A,B,C,D の相対性に基づいた考えである.

図 5.1: 変換テーブル

5.3 Word2Vec

Word2Vec とは, 文脈を用いて単語をベクトル化することで, 単語間の周辺に出現する単語を予測するニューラルネットワークの手法である. 単語をベクトルに変換することで, 入力した単語に対して近いベクトルの単語 (類似した単語) を取得することができる. 類似した単語は類似した文脈で出現することから同義語や類義語は高い類似度を示す. しかし, 反対の意味を表す対義語も類似する文脈で出現することから類似度が高くなる.

5.4 実験データ·条件

調査に用いるデータは, 森本 [1] の作成した変換テーブル, 日本語の単文データ, 英語の単文データである.

表 5.4.1: 使用データ

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

本研究で使用する Word2Vec モデルは,Python のオープンソースライブラリである gensim[5] である. 学習データとして, 日本語単文データと英語単文データを用いて日本語モデルと英語モデルをそれぞれ作成した.vector_size は 1,000,windows は 50 として学習した.

5.5 実験手順

5.5.1 同義語の抽出

1). 変換テーブルから英語訳が一致する ($B_i = D_i$ となる) 日本語単語対 (A_i と C_i) を抽出する.($A_i = C_i$ は除く)

2). 抽出された日本語単語ペア AC それぞれの他の英語訳を変換テーブル中からすべて検索する.

3). 検索された A_i の英語訳と C_i の英語訳を比較し, 完全に一致する日本語単語 AC の組み合わせを抽出する.

手順	A	B	C	D
1	$A_i(\neq D_i)$	$B_i(= D_i)$	$C_i(\neq A_i)$	$D_i(= B_i)$
2	A_i	B_i B_j B_k B_l	C_i	D_i D_n D_o D_p
3	A_i	$B_i(= D_i)$ $B_j(= D_o)$ $B_k(= D_n)$ $B_l(= D_p)$	C_i	$D_i(= B_i)$ $D_n(= B_k)$ $D_o(= B_j)$ $D_p(= B_l)$
	A_i		C_i	

図 5.2: 同義語抽出の手順

5.5.2 類義語の抽出

1). 変換テーブルから英語訳が一致する ($B_i = D_i$ となる) 日本語単語対 (A_i と C_i) を抽出する.($A_i = C_i$ は除く)

2). 抽出された日本語単語ペア AC それぞれの他の英語訳を変換テーブル中からすべて検索する.

3). 検索された A_i の英語訳と C_i の英語訳を比較し, 完全には一致しない日本語単語 AC の組み合わせを抽出する.

手順	A	B	C	D
1	$A_i(\neq D_i)$	$B_i(= D_i)$	$C_i(\neq A_i)$	$D_i(= B_i)$
2	A_i	B_i B_j B_k B_l	C_i	D_i D_n D_o
3	A_i	$B_i(= D_i)$ $B_j(= D_o)$ $B_k(\neq D_i, D_n, D_o)$ B_l	C_i	$D_i(= B_i)$ $D_n(\neq B_i, B_j, B_k, B_l)$ $D_o(= B_j)$
	A_i		C_i	

図 5.3: 類義語抽出の手順

5.5.3 対義語の抽出

1). 変換テーブルから英語訳が一致しない ($B_i \neq D_i$ となる) 日本語単語対 (A と C) を抽出する.

2). 変換テーブル内で対応訳に当たる A_i と B_i(日本語-英語) それぞれについて word2vec を用いて類似度の最も高い日本語単語, 英語単語を抽出する.

3). モデル内の辞書に存在しない単語を含むパターンを除去する.

4). 抽出された類似度の高い日本語-英語単語の組み合わせと同じ対応が A_iB_i ペアに存在すれば対義語と分類する.

手順	A	B	A2	B2	C	D
1	$A_i(\neq D_i)$	$B_i(\neq D_i)$			$C_i(\neq A_i)$	$D_i(\neq B_i)$
2	A_i	B_i	$A(\fallingdotseq A_i)$	$B(\fallingdotseq B_i)$	C_i	D_i
3			A	B	$C_i(\fallingdotseq A)$	$D_i(\fallingdotseq B)$
4	$A_i(\fallingdotseq C_i)$	$B_i(\fallingdotseq D_i)$			$C_i(\fallingdotseq A_i)$	$D_i(\fallingdotseq B_i)$
	A_i				C_i	

図 5.4: 対義語抽出の手順

第6章　実験結果

6.1　同義語

同義語は,93 対抽出できた. 抽出されたデータの一部を以下に示す. また, 例文を示す.

表 6.1.1: 同義語

	C	D	A	B
1	間違い	error,errors mistake,mistakes	ミス	error,errors mistake,mistakes
2	価格	price,prices the prices,	値段	prices,the prices price
3	吸収 する	absorbs	吸い込む	absorbs
4	まかせる	leave	任せる	leave

例文）

1). 彼は自分の**間違い**を認めない. He won't acknowledge his *error*.
その**間違い**を正した. I put the *mistake* right.
私の**ミス**で負けた. We lost because of my *error*.
私の**ミス**をお許しください. Please forgive my *mistake*.

2). 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.

3). 海綿 は 水 を **吸収 する** . A sponge *absorbs* water .
海綿 は 水 を **吸い込む** . A sponge *absorbs* water .

4). その 選択 は 君 に **まかせる** . I *leave* the choice to you .
万事 君 に **任せる** . I *leave* everything to you .

6.2 類義語

類義語は,2,088 対抽出できた. 抽出されたデータの一部を以下に示す. また, 例文を示す.

表 6.2.1: 類義語

	C	D	A	B
1	お昼	lunch,*noon*	正午	at noon, midday,*noon*
2	さかのぼる	dates, *gose* dates back,	行く	I go,go,go to, *gose*,going, going to,to go
3	この 道路	This road, *road*	道	The path,The road, path,*road*,road gose ,road leads,street ,the road,way
4	和平	The peace,peace	平和	Peace,peace peaceful

例文）

1). 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* .
お昼 に は サンドイッチ が 出 た . Sandwiches were served at *lunch* .

2). その 話 は はるか 昔 に さかのぼる . The story *goes* far back into the past .
この 道 は 駅 に 行く . This road *goes* to the station.
この 風習 の 起源 は １２ 世紀 に さかのぼる . The custom *dates* back to the 12th century .

3). この 道路 は １ マイル 先 で 分岐 する . *The road* forks a mile ahead .
道 が あそこ で カーブ する . *The road* curves there .
自動車 に は この 道 は せま すぎる . *The path* is too narrow for the car .

4). 大勢は和平に傾いていた. The general drift of affairs was toward *peace*.
私は平和に暮らしたい. I want to live in *peace*.
平和に事を納める. Bring the affair to a*peaceful* settlement.

6.3 対義語

対義語は,86 対抽出できた. 抽出されたデータの一部を以下に示す. また, 例文を示す.

表 6.3.1: 対義語

	C	D	A	B
1	東	east	西	west
2	総理	prime	大臣	minister
3	年代	1980s	１９７０	1970s
4	太郎	Tarou	花子	Hanako

例文)

1). 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). **総理**は賓客を玄関に出迎えた. The *prime* minister received the guest at the entrance.
大臣が現地を視察した. The *minister* made an on-the-spot inspection.

3). アルゼンチン経済は１９８０**年代**に停滞した. The Argentine economy stagnated in the *1980s*.
この大学は１９７０年代に初めて女子入学を許可した. This college first admitted women is the *1970s*.

4). **太郎**は東京へ行った. *Tarou* went to Tokyo.
textgt 花子はテニスへ行った. *Hanako* played tennnis.

第7章　考察

7.1　評価

実験結果で得られた同義語 · 類義語 · 対義語をランダムに 30 対ずつ選び, 人手と既存辞典 [6][7] で比較した. 同義語辞書は用意できなかったため, 類義語と対義語についてのみ辞典を用いて評価した. 結果を以下に示す.

表 7.1.1: 人手評価と既存辞典 [6][7]

	人手数 (正解率)	辞典数 (正解率)
同義語	17 (56.7%)	
類義語	13 (43.3%)	11 (0.37%)
対義語	7 (23.3%)	6 (20.0%)

人で評価と既存辞典 [6][7] の結果で比較すると人手評価の方が評価数が多い. これは, 既存辞典 [6][7] に記載されていない類義語 · 対義語を提案手法において抽出できたことを示す. 辞書に出現せず, 人手で正解と評価した単語を以下に示す.

表 7.1.2: 類義語 · 対義語出現例

	出力例
類義語	”講演”と”スピーチ”,”芝”と”芝生”
対義語	”太郎”と”花子”

7.2 問題点

7.1 節で単語対が不正解となった原因としては主に 2 つ考えられる.1 つは, 同単語の表記方法の違いである. 変換テーブルにおいて漢字とその読み仮名は別単語として扱われるため, 同じ英語訳を持つ別単語として抽出されたと考えられる.2 つ目は, 対義語抽出において名詞連続複合語の関係にあたる単語対の出現である.Word2Vec では類似する文脈に出現する単語ほど類似度が高くなるが, 連続する名詞は文脈に依存せず同一文で出現できるため, 純粋な文脈の類似度よりも高い類似度を示すと考えられる.

表 7.2.1: 不正解の原因

	読み	複合語	その他
同義語	11	0	2
類義語	7	1	9
対義語	2	8	13

また, 表 7.1.1 より同義語 · 類義語に対して対義語の正解数が少ないことが分かる. これは, 対義語は単語ごとに原則一対なのに対し, 同義語 · 類義語は複数対存在するため正解となる単語が多いことが原因と考えられる.

第8章　今後の課題

8.1　視点の相互性

同義語・類義語抽出の実験手順2において, 変換テーブル中から日本語単語と訳される英単語をすべて検索したが, これは日本語から英語への一方的な視点であり, 検索された英単語には他の日本語単語の訳を持つ可能性がある. 他の訳を持つ場合, 日本語単語同士は英語訳が共通するため類似した単語であると考えられ, 同義語もしくは類義語が研究結果より増える可能性がある. 逆に, 日本語から英語の視点と英語から日本語の視点の翻訳対応が完全に取れている場合, 本研究の結果よりも信頼の高い同義語が抽出できると考える.

8.2　データの数と信頼性

同義語・類義語抽出に関しては翻訳の対応で分類している. しかし, 本来存在するはずの訳が変換テーブル [1] 中に存在しない場合がある. 訳の数が1つ違うだけで, 共通か非共通かが変わるため, 結果に大きく影響する. つまりデータの数が結果に大きく影響するのである. よって, 精度向上にはより多くの対訳文データで変換テーブルを作成し, 同様の実験を行う必要があると考える. また, 変換テーブル [1] は自動生成であり, 精度は95%である. 高い精度をもつ翻訳対応ではあるが, 本研究では誤った訳の一つで結果が左右される. よって, 変換テーブル中で一定の出現頻度があり, 翻訳の可能性が高い単語対のみで実験を行うことで翻訳の間違いを防ぐ必要があると考える.

第9章　おわりに

従来手法では, 単一言語において人手により同義語 · 類義語 · 対義語を分類してきた. しかし, 人手で辞書を作成するのは多大な時間とコストを必要とする. 本研究では,2 言語において変換テーブルを用いて分類の自動化, 辞書作成を行った. 実験結果から, 自動で分類することができたことにより辞書作成の時間を短縮することができた. また, 既存辞典 [6][7] で出現しない同義語 · 類義語 · 対義語を抽出することができた. しかし, 考察と今後の課題より, 精度について改善点があると考える.

謝辞

最後に, 本研究のご指導をいただきました鳥取大学工学部知能情報工学科自然言語処理研究室の村上仁一准教授, 村田真樹教授に厚く御礼申し上げます. そして, 平素よりお力添えいただいた自然言語処理研究室の皆様をはじめ, 参考にさせていただいた論文の著者の方々に深く感謝いたします.

参考文献

[1] 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

[2] 広辞苑　第六版, 岩波書店,2008

[3] 明鏡国語辞典　第二版, 大修館書店,2012

[4] ブリタニカ国際大百科事典 小項目電子辞書版, ブリタニカ・ジャパン,2012

[5] gensim:Topicmodellingforhumans,https://radimrehurek.com/gensim/#

[6] 柴田武 山田進 編, 類語大辞典, 講談社,2002

[7] 中村一男 編, 反対語大辞典, 東京堂出版,1965