定義文から自動獲得した言い換えフレーズペアの分析

河合 剛巨　　橋本 力　　鳥澤 健太郎　　川田 拓也　　佐野 大樹
情報通信研究機構 ユニバーサルコミュニケーション研究所 情報分析研究室

1　はじめに

言い換え知識の獲得は、自然言語処理の様々なアプリケーションにとって有用である。これまでに、Webから自動獲得した定義文より、大量の言い換えフレーズペアが自動獲得できることが示されている [1]。言い換えフレーズペアの自動獲得では互いに類似しているフレーズペアを獲得するため、(1) のように読みが同じフレーズペアや内容語が同一のフレーズペアなどの自明な言い換えも獲得結果には含まれる。

(1) a. 動脈がつまる ⇔ 動脈が詰まる
　　b. 植物や花を育てる ⇔ 花や植物を育てる

一方、(2) は、フレーズとしては言い換えが成立するが述語対 (挙げる, 出す) は同義ではないので、自明ではない言い換えのフレーズペアである。

(2) 収益を挙げる ⇔ 利益を出す

本研究は、このように非自明な言い換えフレーズペアの獲得を目的とする。そのために、Webから自動獲得した定義文より得られたフレーズペアを語彙資源を用いて分析し、膨大な量の自動獲得結果の中から自明ではない有益なフレーズペアを自動検出する。

2　関連研究

言い換えフレーズペアの獲得方法は、分布類似度を用いる方法 [2] と、単言語のパラレルコーパスを用いる方法 [3] の2つに分けることができる。

前者は、大量の (パラレルでない) コーパスを使える利点があるが、出現文脈の分布類似度に依存する方法である。故に、低頻度の表現に対する精度の問題がある。また、同義と反義の関係を区別することが難しい。

後者は、大規模な単言語パラレルコーパスの用意に膨大なコストがかかるという問題がある。橋本法 [1] は後者のアプローチを採用しつつも、Web上の大量の定義文を自動獲得することでこの問題を回避している。この手法はWeb文書6億件より約214万の定義文を獲得している。この中には「捕鯨」や「食物繊維」、「オフ会」等の多様な概念が含まれる。橋本法では、獲得した定義文の集合から同じ概念の定義文を対にすることで約2,966万の定義文対を得ており、この定義文対中に含まれるフレーズペアが言い換えであるか否かを自動判定する。自動判定では、フレーズペア間の類似性と、文脈間の類似性を考慮した素性を用いる。この手法により高精度で言い換えフレーズペアが獲得できるが、中には、読みが同じか、内容語が全て同一のフレーズペアなどの自明な言い換えも多く含まれる。しかし、自動獲得したフレーズペアには、異表記対（ニガウリ，苦うり）や同義語対（赤ちゃん，赤ん坊)、動詞含意関係（完勝する→負かす）などの様々な語彙関係が豊富に含まれている。このような関係を含む言い換えフレーズペアを獲得することは、新たな語彙知識を獲得することにも繋がる。そのためにも、より多くの非自明な言い換えフレーズペアを獲得することが望まれる。

3　分析方法

分析手順は次の通りである。**i)** 自動獲得したフレーズペアに現れる内容語対を、語彙資源中の語対との比較によりアライメントし、対応付けの有無に基づいて4つに大分類する。**ii)** 正しい言い換えや含意関係が含まれると考えられる分類項目 (II, III) の中から、非自明な言い換えや含意関係のフレーズペアを自動検出する。

分類項目は次の4つである。

I. 読みが同じか、内容語が全て同じフレーズペア
II. 全ての内容語が対応付けられるフレーズペア
III. 一部の内容語の対応が取れないフレーズペア
IV. 全ての内容語の対応が取れないフレーズペア

アライメントには、統計的機械翻訳の技術を使う方法 (SMT法) や分布類似度を使う方法、語彙資源を用いる方法がある。SMT法は大量の単言語パラレルコーパスが必要であり、定義文対を使ったとしても低頻度表現に対応するのが難しい。分布類似度を使う方法は、同義や含意などの明確な意味関係と単なる類似の関係を区別するのが難しいという問題がある。フレーズペアは同概念に対する定義文対から得ているため類似性が高いためである。本研究は、語彙資源を用いる方法を取る。語彙資源は、ALAGIN[1]の語彙資源を主とする表1の資源を用いる。

[1]高度言語情報融合フォーラム, http://www.alagin.jp/

図 1: 語彙資源を用いたアライメントの図解

語彙資源を用いたアライメントの手順は次の通りである。各フレーズペアにおいて、体言間の内容語対もしくは用言間の内容語対が、表 1 の語彙資源中に現れていれば、それらを全て対応付ける。また、形態素の原形が同じか、あるいは読みが同じ場合も対応付ける。アライメントの対象とする内容語は名詞、形容詞、動詞である。図 1 でアライメントの補足説明をする。図 1 のフレーズペアの場合、体言間の内容語対 (エックス線,X線) は、読みが同じ関係で対応付けられ、また、語彙資源に同義語対として存在するので、同義語対の関係でも対応付けられる。内容語対 (コンピュータ, コンピューター) は異表記対で対応付けられる。用言間の (使う, 用いる) は動詞含意関係で対応付けられる。

次に、こうした手続きで上記分類項目 I~IV に分類が出来たとして、分類項目 II、III をさらに細分類する (I は自明なので対象外とし、IV は手がかりが少なく高精度の言い換え獲得が期待できないため対象外とする)。

II の「全ての内容語が対応付けられるフレーズペア」については、次の手順で細分類し、言い換えや含意関係のフレーズペアを検出する。**i)** 用言間に動詞含意関係を含むか否かで分ける。**ii)** 動詞含意関係を含む場合には、動詞含意関係の負例[2]を含むペアか否かで分ける。この手順により、II を次の 3 つに細分類できる。

II-(1) 動詞含意関係を含まないフレーズペア
II-(2) 動詞含意関係の負例[2]を含むフレーズペア
II-(3) 動詞含意関係の正例を含むフレーズペア

動詞含意関係の負例は、動詞含意関係データベース構築前の自動獲得では含意を持つ可能性が高かったが人手検証で負例とされた動詞対であり、何らかの意味的関連性はあるものの反例が存在し厳密な意味で含意と見なせない動詞対である。例えば、(制限する, 禁止する) や (保護する, 防ぐ) などである。

II-(1) は同じ用言か同じ読みの用言からなり、II-(3) は用言対が動詞含意関係の正例なので、II-(1),(3) は多くが言い換えや含意関係であると考えられる。II-(2) については、用言を関連語対で絞り込み、例外を除去する。II-(2) は、用言対が動詞含意関係の負例なので、言い換えや含意関係のフレーズペアが獲得できれば、それは非自明なフレーズペアである可能性が高い。

[2]言い換えフレーズペアの一部には成りにくいと考えられるため、動詞含意関係データベースの正例のうち「前提関係」と「作用反作用関係」に属するものは II-(2) の負例として扱う。

表 1: 適用する語彙資源

ALAGIN 日本語異表記対データベース (Ver.1.1)
編集距離 1 の異表記対。人手生成データベース約 6.2 万対、自動獲得データベース (svm.linear.s1) 約 139 万対を用いる。
ALAGIN 基本的意味関係の事例ベース (Ver.1.4)
約 10 万対を異表記対、略記対、同義語対などの意味関係に分類したもの (対義語対と部分・全体語対も含む。これらは余分なアライメントになる可能性もあるが、少数であることこと、否定を伴うと言い換えになる例もあるため、用いることにした。)
基本的意味関係の事例ベースの拡張データ
ALAGIN 文脈類似語データベースの類似度上位のペアを人手でチェックし複数の意味関係に分類した 147 万対。同義語対、異表記対、略記対、誤表記対などの同義関係とみなせるペアと、部分・全体語対、上位下位語対、同類語対などの近い意味関係のペアを含む。部分・全体語対 (骨, 胸骨)、上位下位語対 (名称, 社名) や同類語対 (問題, 悩み) などが使われている場合にも言い換えや含意関係が成り立ちうるため、アライメントに用いる。
ALAGIN 動詞含意関係データベース (Ver.1.3.1)
含意関係のある動詞対を人手によりチェックした約 12.1 万対 (正例:約 5.2 万対, 負例:約 6.9 対)。(事前の実験により、動詞含意関係の負例を含むフレーズペアでも言い換えが成り立つ場合が存在したので、アライメントの段階では全ての関係を用いる。分類結果を細分類する際に、動詞含意関係の正例と負例を区別する。)
推移律により拡張した動詞含意関係
ALAGIN 動詞含意関係データベースのうち、前提関係と作用反作用を除く正例の動詞対に対して推移律 [4] を適用し、3 段までの含意段数で拡張して約 148 万対を得た。これらを、動詞含意関係の正例とみなして用いる。
JUMAN(Ver.6.0) 辞書中の同義・異表記対
辞書に含まれる見出し語の異表記対 約 12 万対

III の「一部の内容語の対応が取れないフレーズペア」は、体言と用言における対応付けの有無により下記 3 つに分け、新たな名詞対の語彙資源獲得ができる III-(1) から言い換えや含意関係のフレーズペアを検出する。

III-(1) 体言だけに未対応の語が残るペア
III-(2) 用言だけに未対応の語が残るペア
III-(3) 体言, 用言ともに未対応の語が残るペア

III-(1) では、動詞含意関係の負例以外の用言対を含むフレーズペアに限定して、橋本法による言い換え判定のスコア[3]で足切りし、例外を除去することで、言い換えや含意関係のフレーズペアを検出する。

また、III については、III-(1)~(3) の区別をしない分析も行った。ここでは、災害や病などに関するものに限定した。III のうち、ALAGIN[1] 負担・トラブル表現リストの負担・トラブル表現を含むペアを抽出し、橋本法によるスコアで足切りし、例外を除去することで、言い換えや含意関係のフレーズペアを検出する。

4 自動獲得したフレーズペアの分析結果

本章では、自動獲得したフレーズペアのアライメントによる分類結果について述べ、分類結果のうち、全ての内容語が対応付けられるフレーズペアと一部の内容語の対応が取れないフレーズペアについて詳細化し、その中から有益なフレーズペアを自動検出しうることを示す。

分析対象は、Web 文書 6 億件から橋本法 [1] で獲得した約 2,966 万対の定義文対より自動獲得したフレーズペ

[3]言い換えか否かの判定に用いた SVM の分離平面からの距離

表 2: 自動獲得したフレーズペアのアライメント内訳

アライメント分類の内訳	ペア数	言い換えフレーズペアの例
I. 読みが同じか, 内容語が全て同じフレーズペア	175,716	くすりをつくる⇔薬を作る ウィルスや細菌を殺す⇔細菌やウィルスを殺す
II. 全ての内容語が対応付けられるフレーズペア (→ 4.2 節で細分類)	722,184	エッセンシャルオイルを用いる⇔精油を使う 決まりを定める⇔ルールを決める
III. 一部の内容語の対応が取れないフレーズペア (→ 4.3 節で細分類)	51,303,356	心臓の**働き**が弱る⇔心臓の**動き**が低下する **マタニティー**生活を送る⇔**妊娠**生活を送る
IV. 全ての内容語の対応が取れないフレーズペア	88,592,489	おしゃべりが出来ない ⇔ 言葉を話せない
計	140,793,745	

アである。このうち、表層が重複するフレーズペアを除外した約 1.4 億対のフレーズペアを用いる。

4.1 フレーズペアのアライメント分類結果

表 2 に語彙資源に基づくフレーズペアのアライメント分類の結果を挙げる。

II の「全ての内容語が対応付けられるフレーズペア」は約 72 万対であり、I よりも十分に多いことが分かる。II には自明でない言い換えや含意関係のフレーズペアが含まれると考えられ、4.2 節で細分類の結果を述べる。

III の「一部の内容語の対応が取れないフレーズペア」は約 5,130 万対あり、II と比べて多い。表 2 の言い換えフレーズペアの例にて**太字**で示した部分が、用いた語彙資源によって対応付けられなかった内容語の例である。このような言い換えフレーズペアが獲得できれば、新たな語彙知識の獲得も可能である。この観点から 4.3 節で細分類する。

IV の「全ての内容語の対応が取れないフレーズペア」は用いた語彙資源ではアライメントの手がかりが得られなかったため対象外とした。しかし、表 2 の例のように意味的に関連性のある語対が含まれるフレーズペアも見られる。

4.2 全内容語が対応付けられるペアの細分類

表 3 に、分類 II の「全ての内容語が対応付けられるフレーズペア」を用言の関係 (動詞含意関係の有無) に基づき 3 つに細分類した結果と、正しい言い換えの割合、正しい含意関係のみの割合を示す。

表 3 より、動詞含意関係を含まない対の言い換えの割合と含意関係の割合の合計は 97%であり、動詞含意関係の正例を含む対の言い換えと含意関係の割合の合計は 95%である。II-(1),(3) からは高精度で言い換えまたは含意関係のフレーズペアが獲得できることが分かる。

動詞含意関係の正例を含む対は (3) のように、体言対は同義 (趣意, 趣旨) か含意関係であり、用言対は動詞含意関係の正例 (したためる→記す) である。つまり、句全体として同義か含意関係の構成要素から成っている。

(3) 趣意をしたためる ⇔ 趣旨を記す

動詞含意関係を含まない対も同様であり、(4) のように用言対は読みが同じか原形が同じ語対などで、体言対は同義か含意関係のある構成要素から構成されている。

(4) 遺体をつつむ ⇔ 遺骸を包む

動詞含意関係の負例を含む対には、(5) のように述語対 (抱える, 負う) は含意ではないがフレーズ全体としては言い換えとみなせるフレーズペアが含まれる。

(5) 債務を抱える ⇔ 債務を負う

さらに、動詞含意関係の負例を含む対を調べたところ、約 99%(366,097 対) は用言に含意、反義、予測関係ではない関連語対 (以下、関連語対) を含むフレーズペアであった。この関連語対を含むフレーズペアについて、次の方法で例外を除外した上で精度を評価した。例外は、冗長なフレーズペアと定義文特有の述語対を有するペアである。冗長なフレーズペアは、各フレーズペアの文末の活用の違いを終止形に揃えた上で、他フレーズペアを文字列包含するペアと、他フレーズペアの読みを包含するペア、他フレーズペアの内容語の一部を同義語に置き換えたペアとする。例えば、ペア「差額から利益を得る⇔差額から利益をあげる」はペア「利益を得る⇔利益をあげる」を文字列包含する。この他に「～の事を言う⇔～の事を指す」のように定義文特有の末尾を有するペアも除外する。

結果を表 4 に示す。この結果から、例外を除去した 79,454 対の 13%が言い換え、56%が含意関係のみのフ

表 3: II. 全ての内容語が対応付けられるペアの分類

内訳	ペア数	割合 (精度) 言い換え	割合 (精度) 含意関係[4]
(1) 動詞含意関係を含まない対	190,851	37%	60%
(2) 動詞含意関係の負例を含む対	370,992	6%	78%
(3) 動詞含意関係の正例を含む対	160,341	18%	77%
計	722,184		

表 4: 関連語対を含むフレーズペアの分類

内訳	ペア数	割合 (精度) 言い換え	割合 (精度) 含意関係[4]
定義文特有の述語対を含む対	74,356	(例外として除外)	
冗長なフレーズペア	212,287	(例外として除外)	
上記例外を除外したペア	79,454	13%	56%
計	366,097		

[4]片方向の含意関係のみの割合。言い換えの割合は含まない。

表 5: III. 一部の内容語の対応が取れないペアの分類

内訳	ペア数
(1) 体言だけに未対応の語が残るペア (→表 6)	20,164,124
(2) 用言だけに未対応の語が残るペア	15,144,032
(3) 体言, 用言ともに部分対応のペア	15,995,200
計	51,303,356

表 6: 体言だけに未対応の語が残るペアの分類

内訳	ペア数	割合 (精度)	
		言い換え	含意関係[4]
a. 一部の体言が未対応のペア	5,585,677	-	-
b. 用言が動詞含意関係負例以外	2,770,484	2%	56%
c. スコアが-0.1 以上のペア	539,509	10%	60%
d. c より例外を除外したペア	188,659	8%	70%

レーズペアであることが分かる。言い換えの精度は上がるが、含意関係の精度が下がるのは例外に含意関係のペアが多かったためと考えられる。

この関連語対を含むフレーズペアからは、例えば、(6) のような言い換えフレーズペアや (7) のような含意関係のフレーズペアが獲得できている。

(6) 紫外線を浴びる ⇔ 紫外線を受ける

(7) 肥満度を判定する → 肥満度をチェックする

4.3 一部の内容語が未対応のペアの細分類

表 5 に、分類 III の「一部の内容語の対応が取れないフレーズペア」を、体言間・用言間の対応関係に基づいて細分類した結果を挙げる。

分類 III-(1) の「体言だけに未対応の語が残るペア」を細分類した結果を表 6 に挙げる。まず、分類 III-(1) のペアのうち、全ての体言が一致しないものを除外する (a)。(a) のうち、用言に動詞含意関係の負例以外を含むペアに限定 (b) しても、言い換えの割合が 2%と少なかった。

そこで、橋本法による言い換え判定スコアが -0.1 以上のペアに限定する (c) ことで、言い換えの割合と含意関係のみの割合の合計は 70%あることが分かる。また、4.2 節と同様に、例外ペアを除外した後で約 18.9 万対のフレーズペアを獲得した (d)。この精度を評価したところ、言い換えフレーズペアが約 8%、含意関係のフレーズペアが約 70% 獲得できることが分かる。

(d) からは、例えば、(8) のような言い換えフレーズペアや、(9) のような含意関係のフレーズペアを獲得した。

(8) 体内の有害物質を外に出す
⇔ 有害物質を体内から排出する

(9) 財産を相続する → 財産の移転を受ける

次に、分類 III の「一部の内容語の対応が取れないフレーズペア」のうち、負担・トラブル表現を含むフレーズペアの分析結果を述べる。表 7 に、細分類した結果を

表 7: III のうち負担・トラブル表現を含むペアの分類

内訳	ペア数	割合 (割合)	
		言い換え	含意関係[4]
a. 負担・トラブル表現を含む	7,403,731	0%	20%
b. スコアが-0.1 以上のペア	234,183	2%	52%
c. b より例外ペアを除外したペア	122,198	10%	40%

示す。分類 III の自動獲得したフレーズペアのうち、負担・トラブル表現を含むものは約 740 万対ある (a) が、言い換えと含意関係のみの精度は高くない。橋本法による言い換え判定スコアが-0.1 以上のペアに限定する (b) と、言い換えの精度は 2%、含意関係のみの精度は 52%である。さらに、4.2 節と同様に例外ペアを除外した後 (c) からは、言い換えフレーズペアが 10%、含意関係のみのフレーズペアが 40%獲得できる。表 6 の (d) ほど含意関係が増えないのは、体言が一致しないペアが含まれているためである。

表 7 の (c) のフレーズペアから獲得した言い換えの例を (10) に、含意関係の例を (11) に挙げる。

(10) 半分の放射能の強さになる⇔放射能が半分減少する

(11) 大地震を起こす → 地震を発生させる

5 結論

我々は、Web から自動獲得した定義文対より得られたフレーズペアに現れる単語対と、既存の語彙資源を用いて、フレーズ間で単語レベルのアライメントを行うことでフレーズペアを分類し、この分類結果に基づいて自明ではない有益なフレーズペアを自動検出できることを示した。今後、自動検出したフレーズペアの人手検証を行い、このフレーズペアを ALAGIN[1] より配信する。また、今後の言い換え研究で活用するとともに、語彙資源を増強する際の手がかりとして利用する予定である。

参考文献

[1] Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ichi Kazama, and Sadao Kurohashi. Extracting paraphrases from definition sentences on the web. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics(ACL2011)*, pp. 1087–1097, 2011.

[2] Dekang Lin and Patrick Pantel. Discovery of inference rules for question-answering. *Natural Language Engineering*, Vol. 7, No. 4, pp. 343–360, 2001.

[3] Regina Barzilay and Kathleen R. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics(ACL2001)*, pp. 50–57, 2001.

[4] 橋本力, 鳥澤健太郎, 黒橋禎夫, 藤田篤, 黒田航, ステイン・デ・サーガ, 村田真樹, 風間淳一. 動詞含意関係データベースの自動拡張. 言語処理学会 第 16 回年次大会 (NLP2010), pp. 940–944, 2010.