構文情報を用いた名詞句の換言

山崎 敦 , 沢井 康孝 , 山本 和英
長岡技術科学大学 電気系
E-mail:{yamazaki,sawai,ykaz}@nlp.nagaokaut.ac.jp

1 はじめに

自然言語では 1 つの意味を表すのに複数の表現が存在するため、表層上異なる表現が同じ意味をもつか判定することが必要である。これを実現する換言処理は自然言語の意味内容を扱う自動要約や質問応答など多くの研究における前処理や文章の簡易化など様々な分野に応用が期待され、研究がさかんに行われている。

現在、シソーラスや様々な研究によって単語同士の換言は容易になっている。しかし、質問応答等での利用には句や文など大きな単位での換言が必要となる。

日本語には単語では換言できないが制約を加えることで換言が可能になる場合がある。例えば、「気持ち」と「念」という対は換言とは言えない。しかし、「感謝の気持ちを込める」と「感謝の念を込める」となれば換言ということができる。つまり、単語での換言はできなくても句での換言は可能である。このように日本語にはある制約のもとで句として換言可能な単語が存在する。本稿では、例に示したようなある制約のもとで句として換言表現を抽出する手法を提案する。

2 関連研究

内容語の置換に重点を置いた「語彙的換言」に関する研究は様々な角度から行われている。

藤田ら [7]、鍛治ら [1] は国語辞典から見出し語と語釈文の対を取り出すことで換言表現を抽出しており、木村ら [2] は単語の表層的な特徴から似ている表現を探し出すことで換言表現の抽出を行っている。既存のシソーラス・類義語辞典などでは、人間が日常的に行っている換言知識を網羅できない。また、単語の表層的な情報でははかれない換言も存在していると考える。

関根 [5] は複数の新聞記事から同じ事柄を報道している記事をコンパラブルコーパスとし、換言表現の抽出を行っている。長谷川ら [6] は大規模コーパスを使い、固有表現との位置関係から換言表現を抽出している。これらは、コンパラブルなコーパスの文を抽出する精度と利用可能なコーパスが限られるという問題がある。また、固有表現という限られたものの位置関係からの抽出手法では取得できる換言表現が狭まれている可能性がある。

本稿では、「同じ使われ方」をしている語には換言の可能性があるという観点から、大規模コーパスから構文情報を用いて換言表現を抽出する手法を提案する。構文情報には係り受け関係を用い、1 節で述べた制約を係り受け関係で得られた係り元と係り先の文節とした。また、係り受け関係から制約を付与し句として換言する対象には名詞が多いと考え、換言の対象を名詞に限定した。

3 提案手法

3.1 手法概要

本手法の処理の流れを図 1 に示す。本手法の処理は前処理部と抽出部の 2 つに分けられる。

- **前処理部**
  まず、構文情報を用いて三つ組の収集を行う。[係り元文節 : 対象文節 : 係り先文節] の組を三つ組とし、構文情報には係り受け関係[1]を用いる。そして、抽出した三つ組の係り元文節と係り先文節の整形を行う。これらの処理を行い、三つ組データとして保持する。
- **抽出部**
  収集した三つ組データの係り元文節と係り先文節から換言候補の抽出を行う。抽出した換言候補について対象文節の整形を行い、共起確率を算出しそのスコアにより妥当性をはかる。

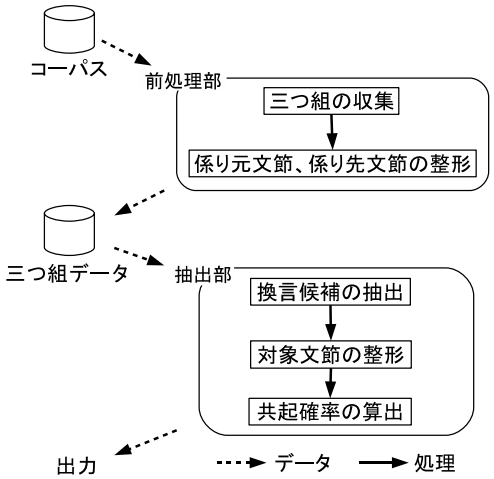

図 1: 処理概要

以下の項では各過程で行われている処理について述べる。

3.2 三つ組の収集

対象となるコーパスから三つ組の収集を行う。まず、コーパスの文を構文解析し、解析結果から [係り元文節 : 対象文節 : 係り先文節] を三つ組として抽出する。

抽出した係り元文節と係り先文節は換言対の制約となる。

本手法は、換言候補を抽出する際、係り元文節と係り先文節をもとにする。よって、係り元文節と係り先文節について対象文節に与える影響が少ない文節は、同義な単語を特定しにくい表現であると考える。そのような文節を含む三つ組は抽出する際に除く。例えば、そのような文節には形容詞や副詞、名詞-副詞可能[2]等が挙げられる。

係り元文節と係り先文節に対して以下の条件に合う文節は除いた。

- 代名詞、形容詞、接続詞、感動詞、副詞、連体詞、フィラー、名詞-固有名詞、名詞-副詞可能で始まる文節。
- 助詞-並立助詞で終わる文節。

[1]本稿では係り受け解析に南瓜 (2) を用いる。
[2]本稿で扱う品詞情報には茶筌 (1) の品詞系を用いる。

・「住所、」のように文節が「名詞 or 未知語 + 読点」の文節。

本手法では名詞句の抽出をタスクとしているため名詞を含んでいるものを対象文節とした。しかし、「同じ使われ方」をもとにしているため、会社名などを対象とした場合、別の会社名を取得してしまう。つまり、固有名詞を対象とした場合、関連語になりやすい。よって、固有名詞は対象文節から除いた。

三つ組収集方法の例を以下に示す。例 1 の文を構文解析すると結果は図 2 のようになる。

例 1) ＮＰＯ等の情報交換の場を提供し、再利用を推進している。

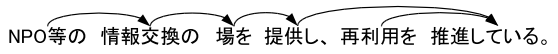

図 2: 例 1 の構文解析結果

「再利用を」を対象文節とした場合、係り元が存在しないため取得することが出来ない。よって、取得できる三つ組は以下に示す 3 つである。

例 2：例 1 から得られる三つ組

2-1. [NPO 等の：情報交換の：場を]
2-2. [情報交換の：場を：提供し、]
2-3. [場を：提供し、：推進している。]

3.3 係り元文節と係り先文節の整形処理

抽出してきた三つ組の係り元と係り先の文節について整形を行う。

3.3.1 係り元文節

係り元文節に関しての整形処理は以下の処理を行う。

- 助動詞と形容詞を削除する。ただし「ない」は削除しない。
- 助詞-終助詞を削除する。
- 動詞を原形に変換する。ただし「ない」が後続する場合は変換しない。

3.3.2 係り先文節

係り先には機能語を含む文節もあるが、それらは制約としては必要ない。例 2-1 の係り先文節は「場を」である。しかし、この文節が「場」であったとしても三つ組で見たときの意味は変化しない。そのような係り先を汎化させ、柔軟なマッチングをするために整形を行う。

係り先文節に関しての整形処理は以下の処理を行う。

- 動詞を原形に変換する。ただし動詞-接尾が後続する場合は変換しない。
- 先頭から順に形態素を見て名詞、動詞、未知語ではない語がきたらその語とそれ以降を削除する。

例 2 を整形すると以下のようになる。「提供し、」の場合、動詞-自立である「し」を原型に変換し、記号である読点が削除され「提供する」となる。「推進している」の場合、「し」を原型に変換し助詞-接続助詞である「て」以降は動詞である「いる」も削除され、「推進する」となる。

3.4 換言候補の抽出

収集した三つ組の「係り元文節」と「係り先文節」を用いて換言候補の抽出を行う。

収集した三つ組の中で係り元文節、係り先文節が同一の三つ組を探し、その 2 つの三つ組の対象文節の対が換言候補となる。

換言候補抽出例を図 3 に示す。例 3 の文に対して、3.2 節及び 3.3 節の処理を行うと例 4 に示す 4 つの三つ組が取得できる。

例 3) あくまでも情報交換の機会を提供する意図で開設したものであります。

例 4：例 3 で得られる三つ組

4-1. [情報交換の：機会を：提供する]
4-2. [機会を：提供する：意図で]
4-3. [提供する：意図で：開設した]
4-4. [意図で：開設した：ものである]

例 2-2 と例 4-1 の係り元文節と係り先文節が同一であるため、その 2 つの対象文節である「場を ⇔ 機会を」という関係が得られる。

3.5 対象文節の整形

3.4 節で得られた換言対について「助詞」と「読点の有無」の同定を行う。日本語は助詞によって取る意味が変化する。例えば「紙の消費」と「減る」の助詞を考えた場合、助詞「を」を取るならば「紙の消費を減らす」、「が」を取るならば「紙の消費が減る」といったように文の意味が変化してしまう。文意が変化するということは換言関係ではなくなるということである。よって、助詞の一致は重要である。

助詞が一致している場合、対象文節の整形を行う。整形処理として助詞を削除する。ただし、助詞-接続助詞の形態素は削除しない。

整形処理の例を図 3 に示す。

3.4 節で得られた関係は、「場を ⇔ 機会を」である。換言候補両者の助詞が「を」で同一であるため整形処理を行う。整形を行うと、換言候補両者の「を」が無くなり「場」、「機会」となる。

よって、係り元文節に「情報交換の」、係り先文節に「提供する」、助詞が「を」の場合に「場 ⇔ 機会」という換言候補を抽出する。

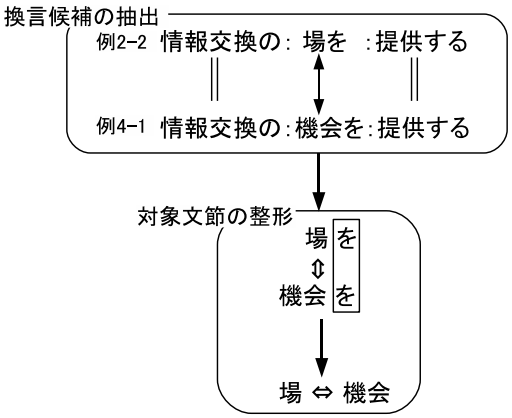

図 3: 換言候補抽出例

3.6 共起確率の算出

三つ組の妥当性をはかるために共起確率を算出する。

式 (1) より係り元文節と係り先文節それぞれからの対象

文節への共起確率としてスコア $C(s_1, s_2, s_3)$ を求める。

$$C(s_1, s_2, s_3) = P(s_1 s_2|s_1) \times P(s_2 s_3|s_3) \quad (1)$$

s_1：係り元文節に出現する文節
s_2：対象文節に出現する文節
s_3：係り先文節に出現する文節

この共起確率に対し閾値を設け、妥当性をはかる。

出現回数が多いということはその換言対の信頼性が高いと考え、閾値は出現回数により異なる数値を設けた。ここでいう出現回数とは換言候補「A ⇔ B」の出現回数である。本手法では換言には三つ組が 2 つ必要でありスコアもそれぞれ付与される。換言に必要な 2 つの三つ組のそれぞれが閾値以上であった場合にのみ換言対として抽出する。

閾値は試行してから設定を行った。出現回数が 1 回のものに関しては正解がほとんどないと考え、本稿では出現回数が 2 回以上のものを対象とした。

4 実験

提案手法を用いて換言対抽出実験を行った。実験に使用したコーパスは我々[4] が作成した Web コーパス約 485 万文 (約 390MB) である。フィルタの閾値を設定するため、各出現回数の精度を求めた。使用したフィルタの閾値とそのときの精度を表 1 に示す。

表 1: 抽出した換言対の各出現回数に対して設定した閾値とそのときの精度及び抽出数

出現回数	閾値	精度 [%]	換言対抽出数
2 回	3×10^{-5}	67	503
3 回	1×10^{-5}	79	366
4 回	3×10^{-6}	65	1180
5 回以上	3×10^{-6}	68	2503

5 結果

実験には抽出した換言対すべてから無作為に 200 個抽出し、独立に人手で評価を行った。

抽出した換言対は 4552 個、精度は 68%であった。

抽出した換言対を大きく 5 つのタイプに分け、各タイプの抽出した比率を表 2 に示す。また、それぞれの結果例も示す。「a ⇔ b : c」の a と b が換言表現、c が出現回数であり、その下に換言を行うための制約を載せている。制約中の □ に換言表現である a と b が入る。例えば、例 5 は「風呂の準備をする」と「風呂の用意をする」で出現回数が 1 回である。例 5 では例示した 2 件の制約の他に換言可能になる制約が 3 件存在するが省略した。

表 2: 抽出した換言対の各タイプの比率

正否	タイプ	比率 [%]
正解 (68%)	A：無条件で換言可能	15
	B：制約付与で換言可能	30
	C：包含関係	16
	D：表記揺れ	7
不正解	E：誤り	32

A. 無条件で換言可能
単語のみで常に換言が可能である関係。
例 5) 準備 ⇔ 用意：5
風呂の：□ を：する
道具の：□ を：する

B. 制約を付与することで換言可能
単語のみでは換言できず、制約を付与し句としての換言が可能な関係。
例 6) 代謝 ⇔ 燃焼：3
脂肪の：□ を：促進する
皮下脂肪の：□ を：促進させる

C. 包含関係
例 7) 生活 ⇔ 日常生活：9
自立する：□ を：送る
住民の：□ に：直結する

D. 表記揺れ
例 8) ご協力 ⇔ 御協力 :12
皆様の：□ を：お願いする
方々の：□ を：頂く

E. 誤り
以下のような誤りが観察された。

・ 反意語
例 9) 支給 ⇔ 支払：3
給与の：□ を：受ける
年金の：□ を：見合わせる

・ 関連語
例 10) 主務大臣 ⇔ 都道府県知事：4
規定による：□ の：認可
規定する：□ の：権限

・ その他
例 11) 方法 ⇔ 問題：5
等の：□ が：ある
ことも：□ と：いえる

B に分類される換言対のように辞典に掲載されていないような換言表現が抽出できた。制約を付与することで換言可能な名詞句が全体の抽出数の 30%、正解の 43%を占めていた。

6 考察

6.1 制約について

制約により、単語の換言が特定できる関係がある。単語のみで見ると「効果 ⇔ 影響」と「効果 ⇔ 威力」という換言対を抽出した。しかし、それらには例 12、例 13 のような異なる制約が得られており、制約により「効果」が適した語に換言できる。有効に制約として働いている例である。

例 12) 効果 ⇔ 影響
及ぼす：□ を：調べる
プラスの：□ を：与える
例 13) 効果 ⇔ 威力
優れる：□ を：発揮する
抜群の：□ を：発揮

同じ制約が複数の換言に存在してしまう問題がある。例えば係り元に「とる」、係り先に「ある」という三つ組で得られる換言対は 10 種類存在した。その一部を例 14 に示す。

例 14) 場合 ⇔ 必要
傾向 ⇔ 場合
傾向 ⇔ 必要
とる：□ が：ある

例 14 に示した単語は句としてみても、それらが換言できるとは言えない。係り元、係り先が制約として弱いものについてはさらにその係り元を見るなどして制約を強くす

べきである。しかし、係り先に関しては制約としてではなく「ある」のような単語のみに係る単語も存在するので一概に削除していいとは限らない。例えば、「可能性 ⇔ 恐れ」という換言対では係り先に「ある」の出現回数が 156 回中 152 回と大半を占めていた。また、「感じ ⇔ 気」の出現回数は 144 回であったが、抽出した係り先文節は全て「する」であった。提案したフィルタはこのような文節を除くためのフィルタであったのだが、コーパスに大量にある文節についてはあまり効果的ではなかった。今後、より多くの事例を観察したうえで、対処法を考えたい。

制約無しで換言可能なものも取れてきている。本手法では制約無しで換言可能なものにも制約を付与してしまっているため、換言できる場所を限定してしまっている。さらに、制約付きでの換言も「係り元、係り先」のいずれかが存在していれば換言可能なものがあった。これらに差異が見つけることができれば、必要な制約のみを付与することができる。単語対の換言知識を用いて、換言対として得られているものを除くことでこの問題は解決できる可能性がある。

対象文節を名詞ではなく名詞句のように繋げて考えれば「重い病 ⇔ 重病」のような句の換言が可能になる。しかし、対象の文節を長くすればするほど係り受け関係の数は減り抽出できるものは少なくなる。さらに、制約の有無や名詞句の判定などが必要となる。

6.2 誤りについて

「同じ使われ方」という観点から処理を行っているため、誤って「反意語」と「関連語」を抽出した。

反意語については字面による判定をすることが出来ないものもあるため、この誤りを削除することは今後の課題である。

関連語の多くは「職業、性別、税」であった。このような単語を対象文節にすることで関連語の辞書を作成することができる可能性もあるが、このような関連語は換言知識には必要ない。三つ組を収集する時点で対象文節が「職業」であるというような知識があればこれらは削除することができる。しかし、これらは固有表現でもなく、見分け方が問題となる。

6.3 抽出数について

今回、出現回数 1 回の関係には触れなかった。出現回数１回には誤りが多いが使用するコーパスサイズによって出現回数は変化する。今回使用した Web コーパスには 485 万文あり、抽出した三つ組は約 660 万組あったのだが、換言できる可能性のある三つ組は約 86 万組まで減少した。今回の抽出基準は換言対の出現回数が 2 回以上というのは三つ組が 4 つ以上必要となるため、抽出できる換言対はさらに少なくなる。

コーパス量による抽出数の変化を確認するため、コーパス量を変化させ、同様の実験を行った。その結果を図 4 に示す。100%が今回実験で使用した 485 万文のコーパス全文である。抽出数が一度大幅に増加してから飽和することが予想されるが、コーパスサイズが 100%でも飽和していないことが読み取れる。そのため、三つ組を大量に集めてくれば換言候補を大量に得られるはずである。本手法では同じ言語であるならば、コーパスの内容に関わらず大量に三つ組を集めることが出来る。

本手法は、係り受け関係を利用している手法なため、日本語だけではなく係り受け関係が存在する言語であれば抽出することが可能だと考える。

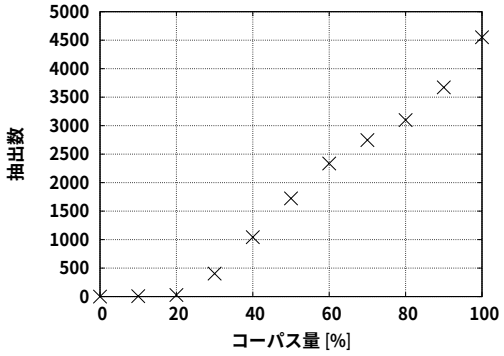

図 4: コーパス量による抽出数の変化

7 おわりに

本稿では、同じ使われ方をしている語に換言の可能性があるのではないかという観点から構文情報を用い、コーパスから換言可能な名詞句を獲得する手法を提案した。Web コーパスを用いた実験の結果、68%の精度で抽出することができた。結果として制約が付与されることで換言可能な名詞句が正解の 43%を占めていた。

制約が無くても換言可能なものや「係り元、係り先」のいずれかが存在していれば換言可能なものについての処理および精度の向上が今後の課題となる。

謝辞

本研究の一部は、科学研究費補助金 若手 (B) 「高密度表現を利用したまとめ型要約に必要な言語変換技術」 課題番号 16700134、及び科学研究費補助金 基盤 (A) 「円滑な情報伝達を支援する言語規格と言語変換技術」課題番号 16200009 によって実施した。

使用した言語資源及びツール

(1) 形態素解析器「茶筌」,Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室,
http://chasen.naist.jp/hiki/ChaSen/

(2) 構文解析器「南瓜」,Ver.0.52, 奈良先端科学技術大学院大学 松本研究室,
http://chasen.org/˜taku/software/cabocha/

参考文献

[1] 鍛治 伸裕, 黒橋 禎夫, 佐藤 理史: 国語辞典に基づく平易文へのパラフレーズ: 情報処理学会 研究報告,NL-144-23,pp. 167-174,2001.

[2] 木村 健司, 徳永 健伸, 田中 穂積: 漢字インデックスを利用したパラフレーズの抽出: 情報処理学会 研究報告,NL-146-7,pp. 39-45,2001.

[3] 佐藤 理史: なぜ言い換え/パラフレーズを研究するのか: 言語処理学会第 7 回年次大会併設ワークショップ, pp. 1-2,2001.

[4] 関口 洋一, 山本 和英: Web コーパスの提案: 情報処理学会 研究報告,NL-157-17,pp. 123-130,2003.

[5] 関根 聡: 複数の新聞を使用した言い替え表現の自動抽出: 言語処理学会第 7 回年次大会併設ワークショップ,pp. 9-14,2001.

[6] 長谷川 隆明, 関根 聡: 教師なし学習による関係抽出に基づくパラフレーズの獲得: 情報処理学会 研究報告,NL-159-27,pp.193-200,2004.

[7] 藤田 篤, 乾 健太郎, 乾 裕子: 名詞言い換えコーパスの作成環境: 電子情報通信学会 技術報告,TL2000-32,2000.