文脈の多様性に基づく名詞換言の提案

梶原 智之　　山本 和英
長岡技術科学大学 電気系
{kajiwara, yamamoto}@jnlp.org

1．はじめに

我々はこれまで、国語辞典の語釈文を用いた内容語の換言について研究してきた[1]。国語辞典の語釈文は、見出し語を平易な数語で説明しているため、見出し語から語釈文中の語への換言により、意味を保持した置換と語彙の平易化が期待される。しかし、語釈文は数語で構成される短文なので、換言候補となる語が少なく、複数の国語辞典を併用するなどの工夫を行っても自然な換言を得ることは難しい。また、語釈文は全体で見出し語と等価であり、語釈文から抽出した各語が見出し語と必ずしも換言可能であるという保証はない。

本稿では、国語辞典など既存の換言知識に頼らず、大規模コーパスから得られる文脈の多様性に基づいた換言を提案する。提案手法では、既存の換言知識を用いる手法と比べて換言候補を多く獲得することが可能であり、その中から換言可能な候補を選択することで、より効果的な換言ができる。

2．提案手法

本稿では、大規模コーパスから得られる文脈の多様性に基づき、文中の名詞を他の名詞に換言する。「似た意味の語は似た文脈で用いられる」という分布仮説[2]に基づき、まず入力文と同じ文脈で用いられる名詞をコーパスから抽出する。そして、抽出した各名詞と入力文中の名詞との、文脈の類似度を格フレーム辞書により計算し、類似度の高い名詞へ換言を行う。

提案手法による名詞の換言の概要を図 1 に示す。本手法は、語の共起頻度を用いず、語の用いられる文脈の種類数のみを用いて換言候補を選択することが特徴である。これは、換言対象の語とより多くの文脈を共有する換言候補の語は、換言可能性がより高いという考えに基づく。

2.1．同じ文脈で用いられる名詞の抽出

本研究では、換言対象の名詞の前後 1 文節を文脈と定義し、入力文と同じ文脈で用いられる名詞をコーパスから抽出する。

まず、入力文を前文脈と後文脈に分け、各々コーパスを探索する。そして、前文脈の後に出現する名詞と後文脈の前に出現する名詞のうち共通する名詞を抽出する。

例えば、「空港へのアクセスを調べる」という入力文に対して、「アクセス」を換言したい場合、「空港への○○」という前文脈と「○○を調べる」という後文脈に分けてコーパスを探索し、○○に該当する名詞のうち共通する名詞を抽出する。図 1 の例では、前文脈と後文脈で共通して用いられる「乗り換え」「料金」「行き方」の 3 単語が抽出される。

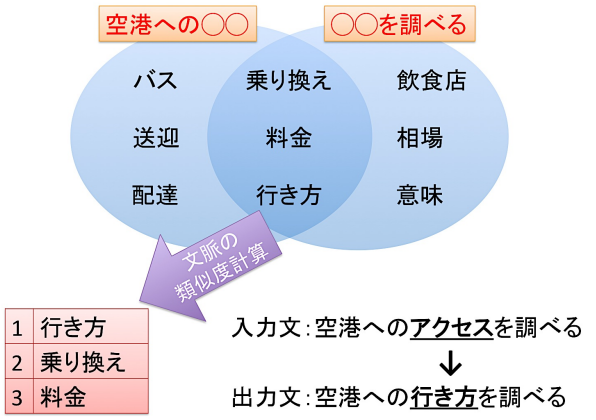

図 1. 提案手法による名詞の換言

表 1. 200 文の換言結果

類似度上位 10 位までに換言可能な語が含まれる	82（41.0%）	類似度 1 位の名詞が換言可能	40（20.0%）
		類似度 2 位から 10 位までの名詞が換言可能	42（21.0%）
類似度上位 10 位までに換言可能な語を含まない	118（59.0%）	同じ文脈で用いられる名詞がない	57（28.5%）
		類似度 10 位までに換言可能な名詞がない	61（30.5%）

2.2. 文脈類似度の計算方法

本稿では、(1) 換言対象の語と換言候補の語が多くの種類の文脈を共有するほど換言可能性は高くなる。(2) 換言候補の語が多くの種類の文脈を持つほど換言可能性は低くなる。という 2 つの仮説を立て、換言対象の名詞と類似した文脈で用いられる名詞を次のスコアが高い名詞と定義する。

$$sim(n_t, n_c) = com(n_t, n_c) * \log(N/DF(n_c)) \quad (1)$$

ただし、n_t は換言対象の名詞、n_c は換言候補の名詞を表し、com は n_t と n_c が共通して用いられる文脈の種類数、N は文脈の総数、DF は名詞 n_c が用いられる文脈の種類数を表す。前項は共通の文脈の種類が多いほど大きくなり、後項は換言候補の文脈が少ないほど大きくなるため、このスコアが高いほど n_t と n_c の文脈が類似していることを表す。

3. 実験方法

3.1. 実験対象

本稿では、Web 日本語 N グラム[GNG]を用いて実験を行った。Web 日本語 N グラムは Web 上の約 200 億文から作成された単語 N グラムで、本稿では最も長い 7 グラムデータを文と見なし、全 570,204,252 文を用いた。これらのうち、先頭が名詞で且つ末尾が動詞の原形である 1,365,705 文を選択し、さらにそのうち頻出する 200 文を抽出して実験対象文とした。この実験対象文のうち、文頭ではない名詞を換言対象の名詞とした。なお、品詞の判別には形態素解析器 MeCab[MEC]を用いた。

3.2. 実験手順

前節で抽出した換言対象の名詞と同じ文脈で用いられる名詞群について、用いられる文脈の類似度を京都大学格フレーム[KCF]を用いて計算した。京都大学格フレームは Web 上の約 16 億文から自動構築[3]された述語とそれが格関係をもつ名詞で、本実験では 34,059 語の述語と 824,639 語の名詞全てを用いた。そして、これらの述語を文脈と仮定し、入力文に含まれる換言対象の名詞を n_t、前節で抽出した名詞群に含まれる各名詞を n_c として(1)式を用いて類似度を計算した。

3.3. 評価

3.1 節で抽出した 200 種類の入力文と換言対象の名詞に対して、3.2 節で計算した類似度の上位 10 位までに含まれる換言候補の名詞を評価した。評価者は、著者のうちの一人である。

4. 実験結果および考察

前章で述べた 200 文に対する換言の結果を表 1 に示す。41%に当たる 82 文は換言可能な名詞を類似度の上位 10 位までに含んでいた。換言不可能な 118 文については、前文脈と後文脈に共通する名詞が存在しない例と類似度 10 位までに換言可能な名詞が存在しない例とが概ね半分ずつを占めた。

次に、換言可能な名詞の出現回数を順位ごとに図 2 に青線で示す。提案手法では、類似度 1 位の名詞が換言可能な例が 82 文中の 40 文と最も多く、本手法による換言の有効性が認められた。なお、換言可能な名詞の順位の平均値は 2.71 であった。

$$sim(n_t, n_c) = com(n_t, n_c) \quad (2)$$

比較のため、(2)式に示すように共通する文脈の種類数のみで類似度を定義した場合の換言可能な名詞の出現回数を順位ごとに図 2 に赤線で示す。比較手法では、換言可能な文の総数は 78 文に減少し、換言可能な名詞の順位の平均値も 3.47 と悪化した。

表 2. 提案手法で換言された例

入力文	出力文
オーナーの**承認**が必要になる	オーナーの**許可**が必要になる
重要な**課題**として取り組んでいる	重要な**問題**として取り組んでいる
良心的な**料金**を提供する	良心的な**価格**を提供する
国内農業の**発展**を阻害する	国内農業の**成長**を阻害する
教育の**拡充**などがあげられる	教育の**強化**などがあげられる

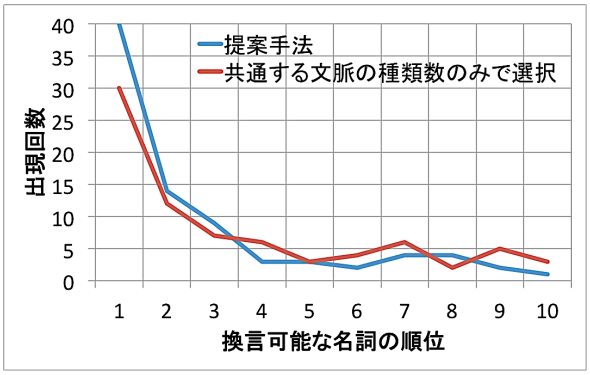

図 2. 換言可能な名詞の順位の出現回数

図 2 からも、1 位から 3 位までの上位では提案手法の出現回数が多く、4 位以下で比較手法の出現回数が多くなっていることがわかる。この結果は、単に多くの文脈で出現する語を除いて、多くの文脈を換言対象の語と共有する語を選ぶことで適切な換言を選択することができることを示している。

表 2 に、提案手法で換言された例を示す。提案手法で出力された名詞は、入力の名詞と共通の文脈を多く持つ名詞であり、置き換えても自然な文で且つ意味を保持できている。

表 3 に、同じ文脈で用いられる名詞がなく換言できなかった例を示す。これらは特に前文脈と強い依存関係にある文が多く、例えば「畜産加工等の○○」という表現は Web 日本語 N グラム中に「畜産加工等の案件」という 1 例しか出現しない。

表 4 に、適切な換言ができなかった例を示す。表 3 の例と同じく、「税抜きの○○」など文脈と強い依存関係がある例が見られた。また、「～～にも○○にも」といった換言対象の語が文脈と並列の関係にある場合は、入力文中の語でなければ意味が保持できない例も見られた。そして、「評価を受ける」

表 3. 同じ文脈で用いられる名詞がない例

畜産加工等の**案件**がある
保全アプローチの**違い**に着目する
降車停留所の**条件**に該当する
更新日順表示に**並び**かえる
家設計プランの**詳細**を見る

表 4. 適切な換言ができなかった例

入力文	換言候補
税抜きの**価格**が混在する	商品、表示
浴衣にも**洋服**にも合う	ドレス、着物、和服…
以上の**評価**を受けている	活動、教育、事業…

のような語単位の換言よりも句単位の換言に適した表現も見られた。例えば「評価を受ける」は「認められる」と句単位で換言するのが適切である。

5. 関連研究

コーパスからの換言獲得としては、パラレルコーパスやコンパラブルコーパスを用いる研究が行われてきた。Barzilay and McKeown[4]は、同じ文書から作られた複数の英訳を用いて換言を獲得している。また、Shinyama and Sekine[5]は、同じ報道を行っている複数の新聞記事を用いて換言を獲得している。これらのパラレルコーパスやコンパラブルコーパスから換言を獲得する手法では、対応する表現同士のアライメントの精度や利用可能なコーパスの量に課題がある。我々の研究では、ノンパラレルコーパスを用いるため利用可能なコーパスの量に制限はない点でこれらの研究とは異なる。

ノンパラレルコーパスから得られる文脈の類似性に基づいて換言を行う研究には、Yamamoto[6]、山崎ら[7]、Marton et al.[8]、Bhagat and Ravichandran[9]などがある。

Yamamotoや山崎らの研究では、コーパスに対する構文解析結果から［内容語：格助詞：内容語］または［係り元文節：対象文節：係り先文節］の三つ組を抽出し、係り受け関係にある表現の共起頻度を計算して換言を獲得している。

Marton et al.の研究では、機械翻訳システムの改良のために未知語の換言を行っている。コーパスから未知語と同じ文脈で出現する単語を換言候補とし、文脈との共起頻度で特徴ベクトルを生成する。そして未知語の特徴ベクトルと各換言候補の特徴ベクトルのコサイン類似度を計算し、最も類似度が高い換言候補へ換言を行うことで機械翻訳システムの精度を改善している。

Bhagat and Ravichandranの研究では、250億語のコーパスから換言を抽出している。コーパス中の単語5グラムを句と見なし、句ごとに相互情報量を用いて特徴ベクトルを生成する。そして同じ文脈を持つ語同士の特徴ベクトルのコサイン類似度を計算し、最も類似度が高い語の組を換言として抽出している。

我々の研究では、単語の出現頻度や共起頻度を計算していない点がこれらの研究と異なる。本稿では文脈の多様性に注目し、語が用いられる文脈の種類数のみを用いて文脈の類似度を計算し、換言先の語を選択する。これは、換言対象の語とより多くの文脈を共有する換言候補の語は、換言可能性がより高いという考えに基づく。

6．おわりに

本稿では、国語辞典など既存の換言知識に頼らず、大規模コーパスから得られる文脈の多様性に基づいた名詞の換言を提案した。提案手法では、入力の文脈に応じた換言が可能であり、換言対象の名詞とより多くの文脈を共有する名詞を換言先に選択するため、適切な換言を得ることができた。

しかし、その網羅性には課題がある。本稿の実験では、Web日本語Nグラムの7グラムデータを文と見なしたり、京都大学格フレームの述語を文脈類似度計算において文脈と見なしたり、項に含まれる名詞のみを扱ったりといった仮定や制限があった。今後は、大規模コーパスの整備を行い、1語対1語の換言だけでなく複数語の換言も視野に入れ、大規模な実験を行いたい。

使用した言語資源およびツール

［GNG］工藤拓, 賀沢秀人. Web日本語Nグラム第1版. 言語資源協会, 2007.
http://www.gsk.or.jp/catalog/gsk2007-c/
［KCF］河原大輔, 黒橋禎夫. 京都大学格フレーム(Ver 1.0). 言語資源協会, 2009.
http://www.gsk.or.jp/catalog/gsk2008-b/
［MEC］工藤拓. MeCab 0.993.
http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

参考文献

［1］梶原智之, 山本和英. 小学生の読解支援に向けた語釈文から語彙的換言を選択する手法. NLP若手の会第8回シンポジウム, 2013.
［2］Zellig S. Harris. Distributional structure. Word, Vol.10, No.23, pp.146-162, 1954.
［3］河原大輔, 黒橋禎夫. 格フレーム辞書の漸次的自動構築. 自然言語処理, Vol.12, No.2, pp.109-131, 2005.
［4］Regina Barzilay, Kathleen R. McKeown. Extracting Paraphrases from a Parallel Corpus. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL), pp.50-57, 2001.
［5］Yusuke Shinyama, Satoshi Sekine. Paraphrase Acquisition for Information Extraction. In Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP), pp.65-71, 2003.
［6］Kazuhide Yamamoto. Acquisition of Lexical Paraphrases from Texts. In Proceedings of the 2nd International Workshop on Computational Terminology (CompuTerm), pp.22-28, 2002.
［7］山崎敦, 沢井康孝, 山本和英. 構文情報を用いた名詞句の換言. 言語処理学会第12回年次大会発表論文集, pp.775-778, 2006.
［8］Yuval Marton, Chris Callison-Burch, Philip Resnik. Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.381-390, 2009.
［9］Rahul Bhagat, Deepak Ravichandran. Large Scale Acquisition of Paraphrases for Learning Surface Patterns. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL), pp.674-682, 2008.