3I4-3

文脈の多様性に基づく名詞換言の評価

Evaluation of noun paraphrasing based on variety of contexts

梶原 智之 Tomoyuki Kajiwara
山本 和英 Kazuhide Yamamoto

長岡技術科学大学 電気系
Department of Electrical Engineering, Nagaoka University of Technology

We paraphrase the noun along the context of the input sentence based on the variety of contexts that is obtained from large-scale corpus. The proposed method has the feature of not using the word frequency nor the co-occurrence frequency but only the number of types of contexts. This is based on the idea that paraphrase candidates appear more common with the target words in the same contexts. For the result of the experiment, this approach was able to produce more appropriate paraphrase than the approaches based on the co-occurrence frequency and PMI.

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[梶原 13]。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。

このような背景から、我々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日本語の名詞換言を提案した[梶原 14]。本稿では、文脈の類似性に基づく他の換言手法との比較から、提案手法の有効性を示す。

2. 関連研究

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としては、Marton et al.が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[Marton 09]。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。また、Bhagat and Ravichandran は 250 億語のコーパスから換言を抽出している[Bhagat 08]。コーパス中の単語 5 グラムを句と見なし、句ごとに自己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。

連絡先：梶原智之，長岡技術科学大学 電気系，新潟県長岡市上富岡町 1603−1，kajiwara@jnlp.org

3. 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[Harris 54]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。図 1 に提案手法による名詞換言の概要を示す。

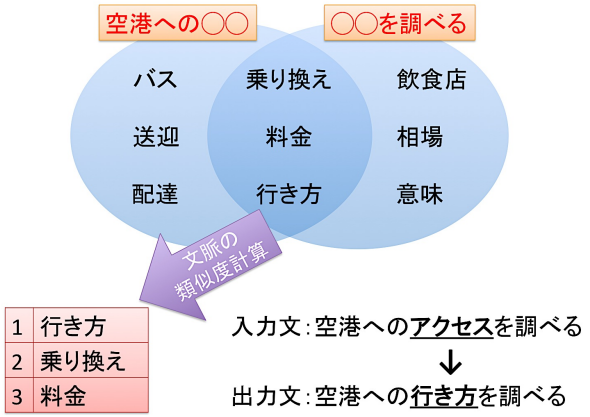

図 1. 提案手法による名詞の換言

3.1 同じ文脈で用いられる名詞の抽出

本手法では、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。

3.2 文脈類似度の計算方法

本稿では、次の 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式(1)の値が大きい名詞と定義する。

(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

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

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。

4. 実験方法

4.1 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。

4.2 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[河原 05]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 として式(1)を用いて文脈の類似度を計算した。

4.3 評価

提案手法を評価するために、関連研究に挙げた文脈の類似性に基づく換言手法との比較を行った。4.1 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、提案手法および Marton et al.の手法[Marton 09]、Bhagat and Ravichandran の手法[Bhagat 08]を用いて類似度の上位 10 位までに含まれる名詞を集めた。評価は、3 人の評価者が換言対象の名詞と入力文中で換言可能な名詞を 1 語ずつ選んだ。

[Marton 09]では、名詞と文脈との共起頻度で名詞の特徴ベクトルを作成し、類似度は特徴ベクトル同士のコサイン類似度で求める。また、[Bhagat 08]では、名詞と文脈との自己相互情報量(PMI)で名詞の特徴ベクトルを作成し、類似度は特徴ベクトル同士のコサイン類似度で求める。両手法とも、Web 日本語 N グラムを用いて、名詞と係り受け関係にある名詞および動詞を文脈と定義し、特徴ベクトルを作成した。式(2)に共起頻度、式(3)に自己相互情報量、式(4)にコサイン類似度をそれぞれ定義する。

$$cooccurrence(w_i, w_j) = \sum_{s_n \in S} freq_n(w_i, w_j) \quad (2)$$

$$pmi(w_i, w_j) = \log\left\{\frac{cooccurrence(w_i, w_j)\sum_{s_n \in S}\sum_{w_m \in s_n} freq_n(w_m)}{\sum_{s_n \in S} freq_n(w_i)\sum_{s_n \in S} freq_n(w_j)}\right\} \quad (3)$$

$$\cos(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}||\vec{v}|} \quad (4)$$

ただし、$s_n \in S$, $w_m \in s_n$, $w_m \in W$ であり、Sは文の集合（コーパス）、Wは単語の集合、$freq_n(w_m)$は文nにおける単語w_mの出現頻度、$freq_n(w_i, w_j)$は文nにおける単語w_iと単語w_jの共起頻度、$\vec{u}, \vec{v}$は特徴ベクトルを表す。

5. 実験結果および考察

前章で述べた 200 文に対する換言の評価結果を図 2 および図 3 に示す。なお、3 人の評価者 A、B、C の kappa 係数は、AB:0.64、BC:0.61、CA:0.59 であり、評価者間の一致度は十分高いと言える。

図 2 は、換言可能と評価された類似度 1 位の名詞数であり、提案手法が 2 つの比較手法よりも多くの名詞を換言できることを示している。

[Marton 09]では多く共起する文脈を重要な文脈と考え、[Bhagat 08]では偏って共起する文脈を重要な文脈と考えている。そのため、[Marton 09]では単体での出現頻度が高い単語が類似度計算に強く反映され、[Bhagat 08]では単体での出現頻度が低い単語が類似度計算に強く反映されている。例えば、[Marton 09]では「こと」が 200 組中 100 組で換言候補として現れており、[Bhagat 08]では「等」「匹」などの接尾辞となる名詞が換言候補として多く現れている。

提案手法は、文脈の出現頻度に依存しないためこれらの影響は少なく、換言対象の名詞と換言可能な名詞のスコアが高くなっている。

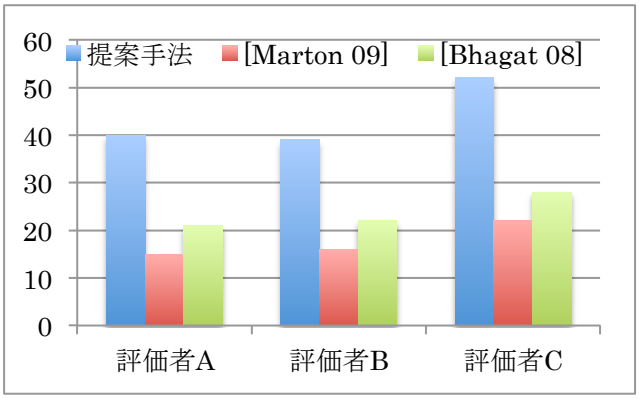

図 2. 換言可能と評価された類似度 1 位の名詞数

図 3 は、換言可能と評価された類似度上位 10 位までの名詞数である。上位 10 位まで見ると、[Bhagat 08]が提案手法の換言可能名詞数に近づいている。ここで、図 4 から図 6 に類似度の順位と換言可能名詞数の関係を示す。提案手法では 1 位の名詞の換言可能数と 2 位の名詞の換言可能数に大きな差があるのに対して、[Bhagat 08]では 1 位から 3 位までの換言可能数の変化が少ない。これは、提案手法では入力文と同じ文脈で用いられる名詞に換言を行うという制限をかけているためだと考える。[Bhagat 08]では入力文の文脈を考慮しないため、入力文において換言可能な語のスコアが最大になる保証はない。例えば、「万円以下の【罰金】に処する」という入力文において【罰金】を換言する場合、[Marton 09]や[Bhagat 08]では「懲役」のスコ

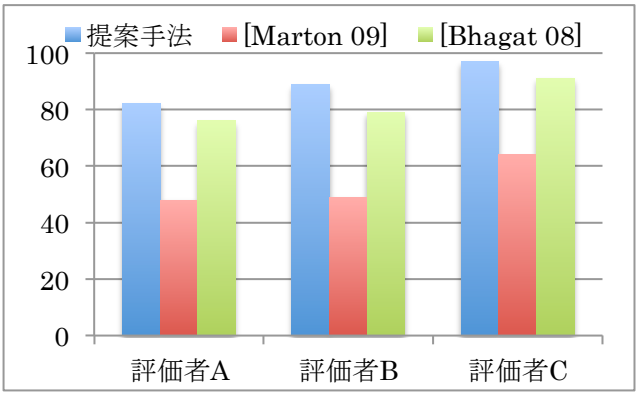

図 3. 換言可能と評価された類似度上位 10 位までの名詞数

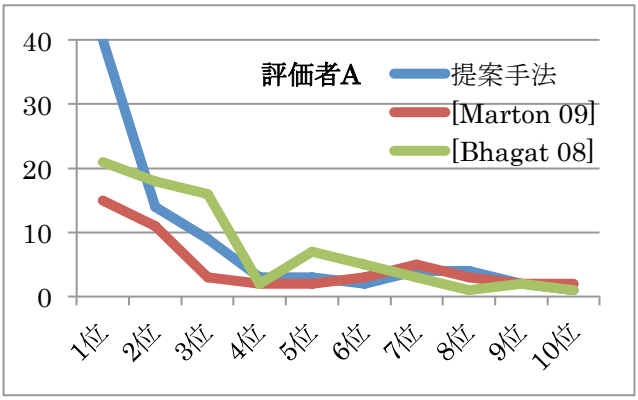

図 4. 類似度の順位と換言可能名詞数の関係（評価者 A）

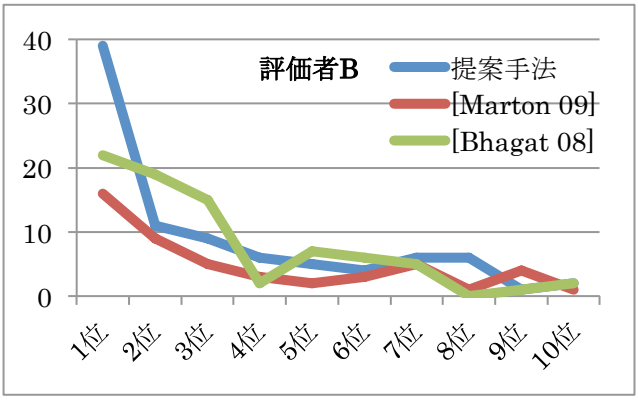

図 5. 類似度の順位と換言可能名詞数の関係（評価者 B）

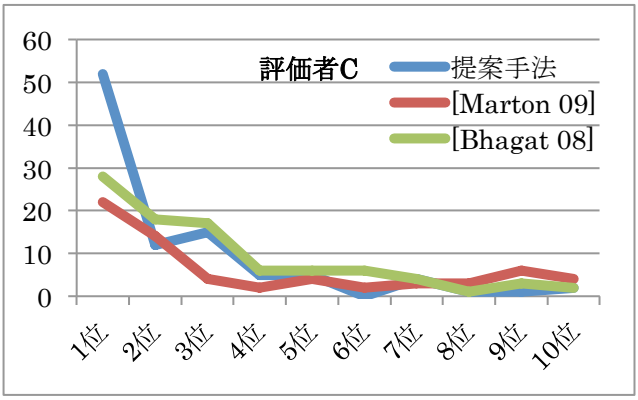

図 6. 類似度の順位と換言可能名詞数の関係（評価者 C）

アが最も高く、次に「科料」や「過料」といった金銭に関する単語が続いている。一方、提案手法では「万円以下の」という入力文の文脈を考慮し、「懲役」という単語は換言候補中に現れておらず、「罰金刑」のスコアが最高で、次に「過料」が続いている。

その他、「腰への【負担】を軽減する」という入力文において【負担】を換言する場合、比較手法では「費用」「経費」「実費」など【負担】の換言先として金銭に関する単語ばかりが換言候補中に出現し、換言可能な語は上位 10 位に存在しない。一方、提案手法では「腰への」という入力文の文脈を考慮し、「負荷」に続いて「ストレス」「ダメージ」「疲労」「緊張」「衝撃」「当たり」「圧迫感」「荷重」「圧迫」と適切な換言候補を挙げることができている。最後に、表 1 に提案手法で換言できた例を挙げる。

表 1. 提案手法で換言できた例

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

6. おわりに

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く名詞の換言手法の有効性を示した。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、単語の出現頻度や共起頻度に関わらず適切な換言を得ることができた。

本研究では 1 語対 1 語の名詞の換言のみを扱ったが、今後は先行研究[梶原 14]でも述べた複数語の換言への拡張を行いたい。

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

[梶原 13] 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, 発表 23, 2013.

[梶原 14] 梶原智之, 山本和英. 文脈の多様性に基づく名詞換言の提案. 言語処理学会第 20 回年次大会発表論文集, D5-1, 2014.

[Marton 09] Y. Marton, C. Callison-Burch and P. 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.

[Bhagat 08] R. Bhagat and D. 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.

[Harris 54] Z. S. Harris. Distributional structure. Word, Vol.10, No.23, pp.146-162, 1954.

[河原 05] 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.