

## 翻訳リペア支援のための Web 日本語 N グラムを用いた 類義語フィルタリング

宮部 真衣<sup>†</sup> 吉野 孝<sup>††</sup>

<sup>†</sup> 和歌山大学大学院システム工学研究科 〒640-8510 和歌山市栄谷 930 番地

<sup>††</sup> 和歌山大学システム工学部 〒640-8510 和歌山市栄谷 930 番地

E-mail: †{s085051,yoshino}@sys.wakayama-u.ac.jp

あらまし 機械翻訳を用いたコミュニケーションにおいて、翻訳リペアは不適切な翻訳箇所を減少させるための方法として重要な役割を果たす。翻訳リペア作業はユーザへの負担が大きいため、修正の必要な単語の類義語や関連語を提示することによる言い換え作業の支援が必要である。しかし、提示数が多い場合、適切なものを選び出すことは容易ではないと考えられる。本稿では、より適切な言い換え候補の抽出のために Web 日本語 N グラムを用いたフィルタリングを提案する。また、2-gram および 3-gram のデータを利用し、前方品詞 2-gram, 後方品詞 2-gram, 3-gram の 3 種類の単語の組み合わせによるフィルタリング実験を行い、以下の知見を得た。(1) Web 日本語 N グラムを用いたフィルタリングにより、90%前後の単語について、言い換え候補を 7 語未満に絞り込むことができ、多数の候補を絞り込むことができる可能性がある。(2) 閾値を 0 とした場合、3-gram によるフィルタリングにおいて抽出失敗率が最も高く (34.4%)、後方品詞 2-gram が最も低く (15.9%) になった。また、除外失敗率については前方品詞 2-gram が最も高く (52.7%)、3-gram が最も低く (28.6%) となった。(3) 複数品詞により構成される言い換え候補については形態素解析を行い、言い換え候補の構成品詞数に応じて利用する N グラムデータを変更することで、抽出失敗を減少できる可能性がある。

キーワード 多言語間コミュニケーション, 機械翻訳, 翻訳リペア, Web 日本語 N グラム

## Synonym Filtering Using Web Japanese N-grams for Translation Repair Support

Mai MIYABE<sup>†</sup> and Takashi YOSHINO<sup>††</sup>

<sup>†</sup> Graduate School of Systems Engineering, Wakayama University 930 Sakaedani, Wakayama, 640-8510 Japan

<sup>††</sup> Faculty of Systems Engineering, Wakayama University 930 Sakaedani, Wakayama, 640-8510 Japan

E-mail: †{s085051,yoshino}@sys.wakayama-u.ac.jp

**Abstract** Translation repair has an important role in creating text that has a minimum amount of improper translations during communications using machine translation. However, deciding the repair strategy to be used is a heavy burden on the user, and a mechanism to reduce the user burden is required. Thus, we will investigate methods to support paraphrasing by providing synonyms and related words for terms that need repair. In this paper, we propose a filter based on Web Japanese N-grams to extract more appropriate paraphrase candidates. In addition, we carried out an experiment using the proposed method, and found the following results. (1) By using a filter based on Web Japanese N-grams, paraphrase candidates were reduced to less than 7 terms in more than 90% of the words. There is a possibility that a large number of candidates can be narrowed down to a few terms. (2) The maximum rate of extraction failure was 34.4% in 3-gram, and the minimum of that was 15.9% in backward part of speech 2-gram at threshold 0. The maximum rate of elimination failure was 52.7% in front part of speech 2-gram, and the minimum of that was 28.6% in 3-gram. (3) Extraction failure is reduced by change to n-gram data corresponding to the number of parts of speech in each paraphrase candidate.

**Key words** multilingual communications, machine translation, translation repair, Web Japanese N-gram

## 1. はじめに

世界規模のインターネットの普及により、ネットワークを介した多言語コミュニケーションの機会が増加している。しかし、一般に多言語を十分に習得することは容易ではない。母国語を用いた多言語話者間のコミュニケーションを実現するために、機械翻訳を利用した取り組みが現在行われている。近年、機械翻訳技術は急速に進展しているが、完璧な翻訳を行うことは困難である。円滑なコミュニケーションを行うためには、不適切な翻訳箇所が少ない文章を作成しなければならない。翻訳文中の不適切な翻訳箇所を減少させるために、入力文章を書き換えていくことを「翻訳リペア」と呼ぶ。翻訳リペアは、機械翻訳を用いたコミュニケーションにおいて重要な役割を果たす。

折り返し翻訳を用いた翻訳リペアにより得られる翻訳結果の精度検証実験から、翻訳リペアによって翻訳精度が改善できることが確認できている [1]。一方、翻訳リペア作業はユーザにとって負担のかかる作業である [2]。そこで、リペア効率改善を目的とした、翻訳の不適切な箇所の強調表示による翻訳リペアへの効果の検証が行われている [3], [4]。しかし、翻訳の難しい箇所を強調したとしても、ユーザは修正内容を自分で考え出さなければならない。しかし、どのように修正するかを考え出すことはユーザにとって負担の大きく、翻訳不適箇所を指摘のみによるユーザのリペア効率の改善は困難であることがわかった [4]。修正の必要な単語の類義語や関連語を提示することによって、ユーザの言い換え作業を支援することができる可能性がある。しかし、単語の類義語や関連語が膨大な場合、適切な言い換え内容を選び出すことは容易ではないと考えられる。本稿では、単語の類義語および関連語を「言い換え候補」と呼ぶ。円滑なリペア作業の支援を行うには、言い換え候補の中から適切なものを抽出する必要がある。

これまでに、類語や同義語の候補から正しい類語・同義語を抽出するための研究が行われている [5], [6]。これらの研究においては、類似度を利用して類語・同義語の候補をランキングし、より適切な類語や同義語の抽出を行っている。本研究では、翻訳リペアにおける言い換え作業の支援が最終的な目的であり、単に言い換える言葉の意味が類似しているだけでなく、対象文中で置き換える際、より適しているものを抽出する必要がある。そこで、Web 日本語 N グラム [7] の利用を検討した。

本稿では、より適切な言い換え候補の抽出のために Web 日本語 N グラムを用いたフィルタリングを提案する。また、提案手法を用いた実験を行い、実験結果についての考察を行う。

## 2. Web 日本語 N グラムを用いた言い換え候補フィルタリング

### 2.1 Web 日本語 N グラム

Web 日本語 N グラム [7] は、言語資源協会が発行している、日本語の単語 n-gram とその出現頻度をまとめた大規模言語リソースである。n-gram データは、2007 年 7 月のスナップショットから構築されており、用いられた総文数は約 200 億文である。このデータは出現頻度が 20 回以上の n-gram を抽出対象に

しており、1~7-gram のデータが「WORD1 [space] WORD2 [space] … WORDN [tab] 出現頻度」のフォーマットで格納されている。

これまでに、複合語の解析 [8] や、新語の意味解析 [9] などへの活用事例がある。

### 2.2 言い換え候補フィルタリング

翻訳リペアにおけるユーザの主な作業は、修正の必要な語句の言い換えである。しかし、言い換え内容を考え出すことは容易ではない [4]。本研究では、言い換える内容として類義語、関連語を提示することによって、リペア作業の効率化を目指す。しかし、言い換え候補となる類義語や関連語は多様であり、膨大な量の言い換え候補が提示される可能性がある。また、もとの単語と言い換え候補を置き換えた場合、不自然な文になる場合もある。このように、言い換えを行う文の文脈に応じて、適切な言い換え候補は異なると考えられる。ユーザにとって多数の候補から適切なものを選択することは時間がかかる作業であると考えられ、できる限り不要な候補を取り除く必要がある。

そこで、不適切な言い換え候補を除外するために、Web 日本語 N グラム [7] を用いたフィルタリングを行う。本研究では、インターネット上の類語辞書<sup>(注1)</sup>により言い換え候補を取得する。原文中の単語のうち、折り返し翻訳文中に存在せず、その言い換え候補も折り返し翻訳文中に存在しないものを「言い換え対象語」とする。言い換え候補の取得およびフィルタリングの手順を以下に示す。

- (1) 機械翻訳により、入力文の折り返し翻訳文を取得する。
- (2) 形態素解析器 MeCab [10] を利用し、入力文および折り返し翻訳文の形態素解析を行う。
- (3) 入力文中の単語 (名詞、動詞) のうち、折り返し翻訳文中に存在しない単語を抽出する。
- (4) 抽出した単語の言い換え候補をインターネット上の類語辞書により検索する。
- (5) 取得した言い換え候補が折り返し翻訳文中に存在しない場合、その類義語のもととなる単語を言い換え対象語とする。
- (6) 入力文中の言い換え対象語とその言い換え候補を置き換える。
- (7) Web 日本語 N グラムを利用し、置き換えた言い換え候補とその前後の単語の組み合わせの出現頻度を求める。
- (8) 出現頻度が設定した閾値以下である場合、言い換え候補から外す。

また、本稿ではフィルタリングにより抽出された言い換え候補数が 7 語以下になることを目標とし、提案手法によるフィルタリングの効果を検証する。なお、抽出数の目標値は、人間が短期的に記憶できる情報量として Miller の論文により発表されたマジックナンバー 7 [11] を参考にし、目安として 7 語に設定した。

## 3. 実験

### 3.1 フィルタリング

Web 日本語 N グラムを用いた言い換え候補フィルタリング

(注1): Yahoo!辞書を利用

表 1 実験に用いたテキストの一部

Table 1 Examples of sentences used in the experiment.

(1) 首相は経済をめぐる諸問題について語った。
(2) ここで私は体積の変化は考えないものとする。
(3) これは汎用機用とパソコン用を合わせたものである。
(4) この写真は兄弟が二人仲良く並んだものです。
(5) 母親というものは自分の子供を弁護するものだ。

これらのテキストは、機械翻訳試験文 [14] のうち、20 文字以上 30 文字以下であるテキストを 100 文選択したものの一部である。

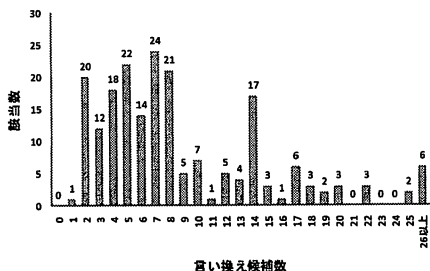


図 1 言い換え対象語の持つ言い換え候補数の分布

Fig. 1 Distribution of the number of paraphrase candidates for each source word.

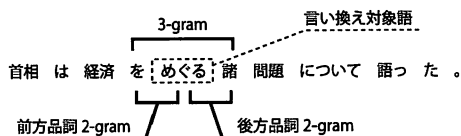


図 2 品詞の組み合わせ

Fig. 2 Example of combination of parts of speech.

の効果を検証するために、単語フィルタリング実験を行った。実験では 2.2 節で述べた手順によりフィルタリングを行う。折り返し翻訳文を取得するための翻訳システムとして、言語グリッド [12] を介して高電社の J-Server [13] を使用した。

実験では、機械翻訳試験文 [14] のうち、20 文字から 30 文字のテキスト 100 文を用いた。これらのテキストは、修正作業が必要だと考えられる文を選択したものである。利用したテキストの一部を表 1 に示す。テキスト 100 文のうち、言い換え対象語が存在した文は 94 文であった。94 文において、言い換え対象語は 200 語、言い換え候補の総数（重複する候補を含む）は 1810 語であった。言い換え対象語の持つ言い換え候補数の分布を図 1 に示す。なお、1 文あたりの言い換え対象語数は平均 2.1 語、最大 5 語、最小 1 語であった。

Web 日本語 N グラムのうち、2-gram（異なり 2-gram 数は約 8 千万）および 3-gram のデータ（異なり 3-gram 数は約 3.9 億）を用いてフィルタリングを行った。フィルタリングの際の品詞の組み合わせは、以下の 3 種類とした。また、それぞれの組み合わせを図 2 に示す。

- 前方品詞 2-gram  
言い換え対象語およびその一つ前の品詞の組み合わせ

表 2 言い換え候補の妥当性の主観評価結果

Table 2 Result of the appropriateness subjective evaluation of paraphrase candidates.

主観評価値	1 以上 2 未満	2 以上 3 未満	3 以上 4 未満	4 以上 5 未満	5	合計
該当数	987	362	266	141	54	1810

- 後方品詞 2-gram  
言い換え対象語およびその一つ後の品詞の組み合わせ
- 3-gram  
言い換え対象語およびその前後の品詞の組み合わせ

前方品詞 2-gram および 3-gram によるフィルタリングに関しては、言い換え対象語が文頭にある場合、単語の組み合わせを作ることができないため、文頭の言い換え対象語を分析対象から除外した。

今回の実験では、検索した単語の組み合わせが N グラムデータ内に存在しない場合を出現頻度 0 とし、出現頻度 0 の場合のみ言い換え候補から外すこととした。

### 3.2 言い換え候補の妥当性評価

フィルタリングが適切に行われたかどうか評価するために、言い換え候補の妥当性に関する主観評価を行った。主観評価では、3 名の評価者により、言い換え対象語と言い換え候補の意味について評価を行う。評価実験の被験者は、和歌山大学大学院の学生 3 名である。

評価においては、以下の 2 組の文を用いる。

- (1) 日本語試験文の原文
- (2) 言い換え対象語と言い換え候補を置き換えた文（言い換え文）

それぞれの文中において、言い換え対象語および言い換え候補を赤字で強調表示し、原文の文脈において、言い換え候補が言い換え対象語と同じ意味で用いられているかどうか比較することにより行う。

評価は Walker らの適合性評価（5 段階評価）[15] により行った。Walker らの適合性評価は、2 名以上の評価者により評価を行うものである。評価基準を以下に示す。

- 5: All (同じ意味)
- 4: Most (文法などに多少問題があるが、大体同じ意味)
- 3: Much (意味は何となく掴める)
- 2: Little (雰囲気は残っているが、もとの意味はわからない)
- 1: None (全く違う意味)

評価は、並べられた 2 文（原文および言い換え文）を見て、1 組の文に対して 30 秒以内で評価するものとした。

表3 フィルタリング結果  
Table 3 Results of filtering.

		フィルタリング適用前	フィルタリング適用後		
			前方品詞 2-gram	後方品詞 2-gram	3-gram
言い換え候補数	平均(語)	9.1	5.0	5.0	3.1
	標準偏差(語)	7.2	6.4	5.3	4.4
	最大(語)	45	38	38	33
	最小(語)	1	0	0	0
フィルタリングが不十分な対象語 * (%)		44.5	16.5	15.0	9.0

\* : 7語以上の言い換え候補をもつ言い換え対象語の割合

表4 フィルタリングの例  
Table 4 Examples of filtering.

(1) あの選手はもうすぐ世界的な水準へ近づく。

言い換え対象語	言い換え候補	主観評価値 *	出現頻度		
			前方品詞 2-gram	後方品詞 2-gram	3-gram
水準	レベル	5.0	389919	40682	2121
	水平	1.7	19955	1201	0
	目安	3.3	46868	401	0

(2) この蝶については、冬、日本に飛来 する ことしかわからなかった。

言い換え対象語	言い換え候補	主観評価値 *	出現頻度		
			前方品詞 2-gram	後方品詞 2-gram	3-gram
する	遣る	1.7	0	2795	0
	為す	2.3	0	20779	0
	行う	3.0	0	7558283	0
	焼く	1.0	0	110622	0
	なさる	3.7	0	54590	0
	執り行う	1.3	0	7799	0
	致す	2.7	0	4856	0
	相嘗む	1.3	0	0	0

\* : 表中の主観評価値は、Walkerらの適合性評価(5段階評価)による評価値である。(5: All, 4: Most, 3: Much, 2: Little, 1: None)

## 4. 実験結果

### 4.1 言い換え候補の妥当性

全言い換え候補 1810 語に関する妥当性の主観評価結果を表 2 に示す。本稿では、平均評価値が「4 以上 5 未満」および「5」であるものを、適切な候補(フィルタリングによって除外されるべきではない単語)とする。また、平均評価値が「1 以上 2 未満」および「2 以上 3 未満」であるものを、不適切な候補(フィルタリングによって除外されるべき単語)とする。なお、「3 以上 4 未満」であるものは、適切な候補と不適切な候補が混在しており、一方に分類することができないため、今回の分析対象

から除外する。

主観評価の結果、総候補数 1810 語のうち、不適切な候補は 74.5%を占めている。これらの候補は言い換えとして用いることができないため、フィルタリングにより除外する必要がある。

### 4.2 フィルタリングによる言い換え候補数の変化

2-gram および 3-gram によるフィルタリング結果を表 3 に示す。フィルタリング適用前の状態では、1つの単語に対する言い換え候補数の最大値は 45 語であった。言い換え候補数の平均は、フィルタリング適用前の状態では 9.1 語であったのに対し、フィルタリング後は 5.0 語(前方品詞 2-gram, 後方品詞 2-gram) および 3.1 語(3-gram) まで絞り込むことができている。また、言い換え対象語 200 語のうち、7語以上の言い換え候補を持つ言い換え対象語の割合は、フィルタリング前は 44.5%を占めている。フィルタリング後(前方品詞 2-gram, 後方品詞 2-gram, 3-gram) はそれぞれ 16.5%, 15.0%および 9.0%に減少しており、9割前後の言い換え対象語について、言い換え候補を 7語未満にすることができている。

### 4.3 フィルタリング例

フィルタリングの例を表 4 に示す。例(1)では、3語の言い換え候補が存在するが、前方品詞 2-gram および後方品詞 2-gram では絞り込みが行われず 3語のままであり、3-gram では 1語へと絞り込みを行っている。この例では、前方品詞 2-gram においては、「水準」の前の語である「な」(Mecabによる形態素解析結果:助動詞)と言い換え候補を組み合わせることで出現頻度を求め、後方品詞 2-gram においては、言い換え候補と、「水準」の後の語である「へ」を組み合わせることで出現頻度を求めている。その結果、前方品詞 2-gram, 後方品詞 2-gram のどちらにおいても、全ての言い換え候補の出現頻度が 0より大きいため、言い換え候補の除外が行われていない。一方、3-gram においては「水準」の前後の語句と、言い換え候補の組み合わせにより、出現頻度が 0である「水平」および「目安」を言い換え候補から除外している。また、主観評価により適切な候補であると評価されたもの(主観評価値 4以上)は、「レベル」という候補であり、3-gram によるフィルタリングでは、適切に絞り込みが行われている。

また、例(2)では、「する」という語の言い換え候補として、8語が存在している。後方品詞 2-gram では、「相嘗む」という候補のみ除外されている。しかし、このテキストでは、どの言い換え候補も文脈的に不自然であり、主観評価値が 4以上である候補は存在しない。後方品詞 2-gram では、後の語である「こと」との組み合わせのみで出現頻度を求めたため、絞り込みができていなかった。一方、前方品詞 2-gram および 3-gram では、前の語である「飛来」と言い換え候補を組み合わせることで出現頻度を求めており、どの語についても言い換え候補から除外できている。

### 4.4 フィルタリングの精度

フィルタリングを行った際の失敗として、以下の 2点について検証を行う。

- (1) 適切な候補を除外してしまう

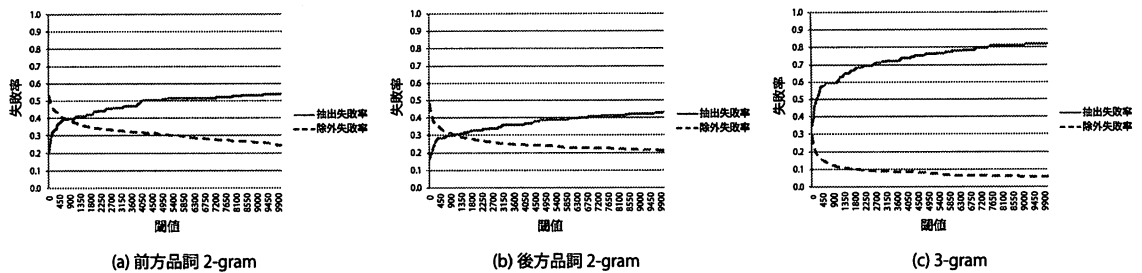


図 3 抽出失敗率と除外失敗率の推移

Fig. 3 Transition rate of extraction failure and elimination failure.

表 5 閾値 0 における抽出失敗率

Table 5 Rate of extraction failure at threshold 0.

	前方品詞 2-gram*	後方品詞 2-gram	3-gram*
適切な候補の総数 (語)	180	195	180
出現頻度 0 である適切な候補の数 (語)	38	31	62
抽出失敗率 (%)	21.1	15.9	34.4

※<sup>1</sup> 総言い換え候補数: 1810 語

※<sup>2</sup> 前方品詞 2-gram および 3-gram によるフィルタリングに関しては、言い換え対象語が文頭にある場合、単語の組み合わせを作ることができないため、文頭の言い換え対象語を分析対象から除外した。

表 6 閾値 0 における除外失敗率

Table 6 Rate of elimination failure at threshold 0.

	前方品詞 2-gram*	後方品詞 2-gram	3-gram*
不適切な候補の総数 (語)	1297	1349	1297
出現頻度 0 以上である不適切な候補の数 (語)	683	644	371
除外失敗率 (%)	52.7	47.7	28.6

※<sup>1</sup> 総言い換え候補数: 1810 語

※<sup>2</sup> 前方品詞 2-gram および 3-gram によるフィルタリングに関しては、言い換え対象語が文頭にある場合、単語の組み合わせを作ることができないため、文頭の言い換え対象語を分析対象から除外した。

## (2) 不適切な候補を除外できない

本稿では、適切な候補を除外してしまう割合を「抽出失敗率」、不適切な候補を除外できない割合を「除外失敗率」と呼ぶ。閾値を 0 とした場合の、抽出失敗率を表 5 に、除外失敗率を表 6 に示す。また、閾値を変化させた場合の各失敗率について、前方品詞 2-gram、後方品詞 2-gram および 3-gram の推移を図 3 に示す。閾値が小さいほど、言い換え候補の除外率は低くなる。すなわち、閾値が 0 に近いほど、抽出失敗率は低くなり、閾値が 0 の場合が最も低い値となる。一方、除外失敗率は、閾値が大きくなるほど低くなる。

閾値が 0 の場合、3-gram の抽出失敗率が最も高く、34.4% の候補の抽出に失敗している。一方、3-gram の除外失敗率は

28.6% となっており、今回用いた組み合わせにおいては、より多くの不適切な候補が除外できている。前方品詞 2-gram および後方品詞 2-gram については、抽出失敗率がそれぞれ 21.1%、15.0% と低いものの、除外失敗率はそれぞれ 52.7%、27.7% となっており、不適切な候補の半数程度を除外することができない。

## 5. 考察

### 5.1 抽出失敗率と除外失敗率の推移

図 3 より、3-gram は 2-gram では閾値による推移が大きく異なる。閾値を変化させた場合、3-gram においては、閾値 1850 を超えた段階で除外失敗率が 10% 以下となるが、抽出失敗率は 60% を超えており、抽出失敗率と除外失敗率の差が大きく開いている。一方、前方品詞 2-gram および後方品詞 2-gram において、除外失敗率が 10% 以下となる閾値は、それぞれ 109150、142850 となっており、閾値が小さい段階では抽出失敗率と除外失敗率の差はあまり開いていない。

フィルタリングを適用する場合、フィルタリングの適用時の状況に応じ、言い換え候補数が多い場合は除外失敗率の低い N グラムデータの利用や大きな閾値の設定、また言い換え候補が少ない場合は、抽出失敗率の低い N グラムデータの利用や、閾値を小さく設定するなど、利用データや閾値の設定を変更する必要があると考えられる。

なお、これらの結果は今回の実験で用いたテキストによる限定的なものであり、今後別のテキストを用いて結果の妥当性を検証する必要がある。

### 5.2 抽出失敗の原因

フィルタの失敗のうち、除外失敗率は閾値を大きくすることで下げることができる。一方、抽出失敗率は閾値 0 の場合よりも下げることができない。そこで、出現頻度が 0 である適切な候補について、抽出失敗の原因を分析した。結果を表 7 に示す。

今回用いたデータにおいては、(A) 単にその組み合わせ自体が存在しない、(B) 言い換え候補が 2 品詞以上で構成されている(「マジックペン」(マジック+ペン) や「その時」(その+時) など) という 2 種類に分類できた。(A) の言い換え候補については、N グラムデータ中には存在しない組み合わせのため、提案するフィルタリングでは抽出することができない。(B) の言い換え候補については、フィルタリングの際、言い換え候補

表 7 出現頻度 0 であった適切な候補の抽出失敗原因

Table 7 Reason for the extraction failure of appropriate candidates.

	前方品詞 2-gram	後方品詞 2-gram	3-gram
(A) 組み合わせが存在しない	17	5	41
(B) 言い換え候補が 2 品詞以上で構成されている	21	26	21

※出現頻度 0 であった、適切な候補数は、前方品詞 2-gram : 28 語, 後方品詞 2-gram : 31 語, 3-gram : 62 語である。

を形態素解析せず、1 語として扱っていることが原因で失敗している可能性がある。そのため、言い換え候補と組み合わせる品詞数 (2-gram では 1, 3-gram では 2) と、言い換え候補の構成品詞数とを合わせた数の N グラムデータを検索することで、本当にその組み合わせが存在しないのかを検証することができる。例えば、「外科手術」という言い換え候補については、「外科」「手術」の 2 品詞に分解できるため、前方品詞 2-gram でフィルタリングを行っている場合、この言い換え候補については前方の 1 品詞 + 言い換え候補 2 品詞とし、3-gram のデータにより出現頻度を検索することが可能である。

(B) の言い換え候補について、上記の手順で出現頻度が存在するか検証した結果、前方品詞 2-gram については 13 候補 (62%)、後方品詞 2-gram については 23 候補 (88%)、3-gram については 9 候補 (43%) が、出現頻度が 0 以上であった。したがって、言い換え候補についても形態素解析を行い、その品詞数に応じて利用する N グラムデータを変更することにより、抽出の失敗を減少させることができると考えられる。

## 6. おわりに

翻訳リペアにおける主な作業は、修正の必要な語句の言い換え作業である。我々は修正の必要な単語の類義語や関連語を提示することによる、言い換え作業の支援に着目し、より適切な言い換え候補の抽出のための Web 日本語 N グラム [7] を用いたフィルタリングを提案した。また、フィルタリングの効果を検証するために、2-gram および 3-gram のデータを利用し、前方品詞 2-gram、後方品詞 2-gram、3-gram の 3 種類の単語の組み合わせによるフィルタリング実験を行った。また、フィルタリング精度の検証のため、語句の妥当性に関する主観評価を行った。

実験の結果、以下のことがわかった。

- (1) Web 日本語 N グラムを用いたフィルタリングにより、90%前後の単語について、言い換え候補を 7 語未満に絞り込むことができおり、多数の候補を絞り込むことができる可能性がある。
- (2) 閾値を 0 とした場合、3-gram によるフィルタリングにおいて抽出失敗率が最も高く (34.4%)、後方品詞 2-gram が最も低く (15.9%) になった。また、除外失敗率については前方品詞 2-gram が最も高く (52.7%)、3-gram が最も低く (28.6%) となった。

- (3) 複数品詞により構成される言い換え候補を、形態素解析せずに 1 語として扱っていたことが抽出失敗の一因であり、言い換え候補の構成品詞数に応じて利用する N グラムデータを変更することで、抽出失敗を減少できる可能性がある。

3-gram は比較的閾値が小さい段階で抽出失敗率が高くなり、抽出失敗率と除外失敗率の差が大きくなるが、2-gram については閾値が小さい段階では抽出失敗率と除外失敗率の差は比較的小さいなど、利用するデータによって各失敗率の推移が異なっている。そのため、フィルタリングの適用時の状況に応じ、言い換え候補数が多い場合は除外失敗率の低い N グラムデータの利用や大きな閾値の設定、また言い換え候補が少ない場合は、抽出失敗率の低い N グラムデータの利用や、閾値を小さく設定するなど、利用データや閾値の設定を変更する必要があると考えられる。

今後は、今回の実験で用いていないテキストデータを用いたフィルタリング実験を行い、本稿の実験結果の妥当性に関する検証を行う。

謝辞 本研究の一部は、日本学術振興会科学研究費 基盤研究 (B)(19300036) の補助を受けた。

## 文 献

- [1] 宮部真衣 他, “折返し翻訳を用いた翻訳リペアの効果,” 電子情報通信学会論文誌, Vol.J-90-D-I, No.12, pp.3142-3150, 2007.
- [2] 小倉健太郎 他, “目的指向の異言語間コミュニケーションにおける機械翻訳の有効性の分析: 異文化コラボレーション ICE2002 実証実験から,” 第 65 回情報処理学会全国大会論文集, 2003.
- [3] 林田尚子 他, “翻訳エージェントによる自己主導型リペア支援の性能予測,” 電子情報通信学会論文誌, Vol.J88-D-I, No.9, pp.1459-1466, Sep.2005.
- [4] 宮部真衣 他, “翻訳不適箇所指摘による翻訳リペアコストと翻訳リペア精度の評価, 情報処理学会デジタルドキュメント研究会, DD-63, pp.29-36, 2007.
- [5] 王玉馨 他, “単語類似度ネットワークを通じた自動同義語獲得,” 情報処理学会自然言語処理研究会, NL-185, pp.7-14, 2008.
- [6] 相澤彰子, “大規模テキストコーパスを用いた語の類似度計算に関する考察,” 情報処理学会論文誌, Vol.49, No.3, pp.1426-1436, 2008.
- [7] 工藤拓, 賀沢秀人著, “Web 日本語 N グラム第 1 版”, 言語資源協会発行, 2007.
- [8] 内山清子, “Web 日本語 N グラムの活用事例,” GSK セミナー Web 日本語 N グラムの利用, 講演資料集 (2008-10-31).
- [9] 荒牧英治, “Google n-gram データによる新語の意味解析,” GSK セミナー Web 日本語 N グラムの利用, 講演資料集 (2008-10-31).
- [10] McCab, <http://mecab.sourceforge.jp/>
- [11] George A. Miller, “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information,” The Psychological Review, vol.63, pp.81-97, 1956.
- [12] Toru Ishida, “Language Grid: An Infrastructure for Intercultural Collaboration,” IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06), pp.96-100, 2006.
- [13] KODENSHA, <http://www.kodensha.jp/>
- [14] NTT Natural Language Research Group, <http://www.kecl.ntt.co.jp/icl/mtg/resources/index.php>
- [15] Kevin Walker, et al., “Multiple-Translation Arabic (MTA) Part 1,” Linguistic Data Consortium (LDC) catalog number LDC2003T18 and ISBN 1-58563-276-7.