複数の機械翻訳を用いた言い換え認識の評価用コーパス構築に向けて

鈴木 由衣　　梶原 智之　　小町 守
首都大学東京
{suzuki-yui, kajiwara-tomoyuki}@ed.tmu.ac.jp, komachi@tmu.ac.jp

1　はじめに

本研究では、日本語の言い換え認識タスクの評価のための単言語パラレルコーパスを構築することを目的に、複数の機械翻訳を用いて言い換え候補を収集する。

同じ意味を表す異なる表現を言い換えと言う。例えば、情報検索や質問応答の際には、ユーザが入力する多様なクエリに対して柔軟な照合が要求されるため、言い換え認識によって表現の多様性を吸収することが重要である。また、子どもや言語学習者のための文章読解支援や文章執筆支援として、入力文の意味を保持したまま平易な表現や流暢な表現へ変換する言い換え生成も活発に研究されている。

このように、言い換え技術は多くの自然言語処理応用タスクの性能改善のために有用であるが、言い換え技術そのものの開発や評価のためのコーパスは少ない。英語では、Microsoft Research Paraphrase Corpus (MSRP) [1] という言い換え認識の評価用コーパスが存在する。しかし、日本語では言い換え技術の開発や評価を目的として構築されたコーパスは存在しない。

そこで本研究では、日本語の言い換え認識タスクに焦点を当て、その評価のために利用可能な単言語パラレルコーパスを構築する。本研究の概要を図 1 に示す。我々は複数の翻訳器を用いて同じ英語文の日本語訳を複数個得た。翻訳が成功しているとき、これらの複数の日本語訳は同じ意味を表す異なる表現だと考えることができ、言い換え候補とする。このようにして得られた言い換え候補を人手で確認し、我々は 363 文対の正例と 102 文対の負例からなる 465 文対の日本語の言い換えコーパスを構築した。ここで、非文はコーパスに採用しなかったが、アノテータは日本語訳のみを見るため機械翻訳の妥当性については確認していない。そのため、言い換え候補の中には流暢な誤訳が含まれている可能性がある。流暢かつ妥当な翻訳は言い換えの正例、流暢な誤訳は言い換えの負例となるので、これはコーパスにバランス良く正例と負例を混ぜることを助けると期待できる。また、本研究では言い換え候補を、それらの 2 文間の単語一致率によって均等にサンプリングした。これによって、単語一致率の高い自明な言い換え事例だけではなく、単語一致率の低い非自明な言い換え事例を積極的に収集した。

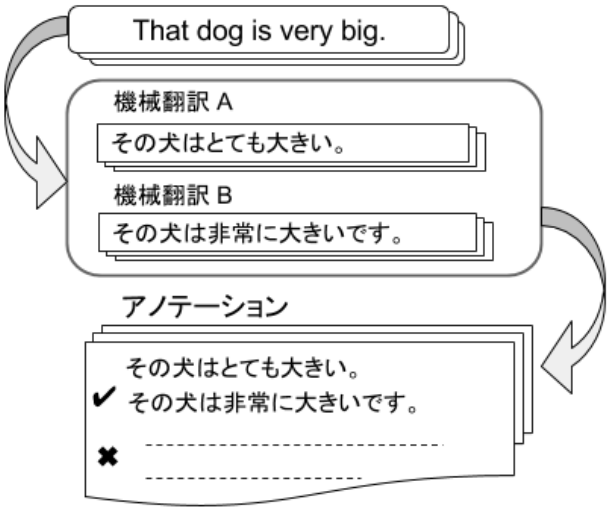

図 1: 複数の機械翻訳を用いた言い換え文対の収集

本研究の主な貢献は、以下の 2 点である。

- 単語一致率の低い非自明な言い換え事例を積極的に収集した。
- 複数の機械翻訳を用いることで言い換え候補の収集コストを抑えた。

2　関連研究

MSRP [1] は言い換え認識タスクの標準的[1]な評価用コーパスであり、3,900 文対の正例と 1,901 文対の負例を含む 5,801 文対からなる。この言い換えコーパスは、ニュース記事から編集距離などのヒューリスティックによって収集された 49,375 文対に対して、文字列の類似度などを素性とする 2 値分類器（Support Vector Machine）によって 5,801 文対の言い換え候補を自動的に抽出し、最終的に 3 人のアノテータが多数決によって正例および負例の言い換えラベルを付与したものである。ヒューリスティックと分類器を用いた言い

[1] https://aclweb.org/aclwiki/index.php?title=Paraphrase_Identification_(State_of_the_art)

換え候補の自動抽出によって、人手評価のコストを抑えるとともに、正例に近い負例を収集することを狙っている。しかし、藤田ら [2] が指摘しているように、編集距離の小さい文対を候補にするなどのヒューリスティックはカバーできる言い換えの種類を限定してしまうという課題を抱えている。この課題を解決するため、本研究では、複数の翻訳器が生成した文対から単語一致率によって均等に言い換え候補をサンプリングすることで、単語一致率の低い（すなわち編集距離の大きい）非自明な言い換え事例を積極的に収集する。

言い換え文対を収集する研究は、これまでも盛んに行われてきた。例えば、あるテキストに対する複数の人手翻訳 [3, 4] や、動画 [5] や画像 [6] に対する複数の説明文から、言い換え文対が収集されている。人手で文を生成しているこれらの先行研究とは異なり、本研究では複数の機械翻訳を用いて自動的に文を生成することによって、言い換え候補の収集コストを抑える。

日本語では、語彙平易化の評価用データ [7, 8] の中に、言い換え事例が含まれている。これらは、どちらも約 2,000 文のコーパスであり、各文中の 1 単語について複数の語彙的な言い換えが 5 人のアノテータによって付与されている。これらのコーパスは内容語の言い換えのみに焦点を当てているが、本研究では種々の言い換えを含む文単位の言い換え対を収集する。また、テキスト間含意関係認識の評価用データ [9] の中にも、両方向の含意関係にある 70 文対の言い換え事例が含まれている。これらは種々の言い換えを含む文単位の言い換え対ではあるが小規模のため、本研究ではより大規模に日本語の言い換え文対を収集する。

3 複数の機械翻訳を用いた言い換え文対の収集

本研究では、2 つの機械翻訳を用いて入力文に対して 2 種類の日本語訳（言い換え候補）を得る。これらの日本語訳には翻訳誤りなどの理由で言い換え候補として不適切な文対が含まれるため、Quality Estimation によって尤もらしい言い換え候補のみを選択する。得られた全ての言い換え候補に対して、アノテータが人手で正例または負例の言い換えラベルを付与する。

我々は Google Translate[2]の PBMT[3]および NMT を用いて、English Wikipedia[4]から抽出した英文について、それぞれ 2 種類の日本語訳を得た。ここで、翻訳誤りを避けるために、言語モデル確率の高い上位 50 万文の英文のみを翻訳した。この言語モデルは、KenLM[5]を用いて English Gigaword Fifth Edition (LDC2011T07) から 5-gram 言語モデルを構築した。このようにして得た日本語訳の組に対して、著者の 2 人が正例または負例の言い換えラベルを人手で付与した。ただし、翻訳誤りを避けるために、式 1 で定義する翻訳品質の高い順に 2,000 文対を対象とした。

$$\begin{aligned} QE_i = \mathrm{BLEU}(e_i, \mathrm{PBMT}_{je}(\mathrm{PBMT}_{ej}(e_i))) \\ \times \mathrm{BLEU}(e_i, \mathrm{NMT}_{je}(\mathrm{NMT}_{ej}(e_i))) \end{aligned} \quad (1)$$

ここで、e_i は i 番目の英文、PBMT_{je} は PBMT を用いた日英翻訳、PBMT_{ej} は PBMT を用いた英日翻訳、NMT_{je} は NMT を用いた日英翻訳、NMT_{ej} は NMT を用いた英日翻訳、$\mathrm{BLEU}(x, y)$ は文 x と文 y の文単位の BLEU スコア [10] を意味する。この翻訳品質が高いというのは、いずれの機械翻訳においても翻訳の前後で意味的な差異が少ないということを表す。WMT2016 の Quality Estimation Shared Task で最高性能を達成した YSDA [11] でも、入力文の言語モデル確率や入力文と折り返し翻訳との BLEU が特に有効な素性であることが示されている。

藤田ら [2] が指摘しているように、MSRP などの言い換えコーパス構築の先行研究では、単語一致率の高い自明な正例が多い。このような特徴を持つ言い換え認識の評価用コーパスでは、表層的な手掛かりのみで問題がある程度解けてしまうという課題がある。そこで本研究では、人手で正例または負例のラベル付けを行う 2,000 文対の言い換え候補を、式 2 に示す単語一致率によって均等に 200 文対ずつサンプリングすることによって、単語一致率の低い非自明な言い換え事例を積極的に収集した。

$$Jaccard(j_i^{\mathrm{PBMT}}, j_i^{\mathrm{NMT}}) = \left| \frac{j_i^{\mathrm{PBMT}} \cap j_i^{\mathrm{NMT}}}{j_i^{\mathrm{PBMT}} \cup j_i^{\mathrm{NMT}}} \right| \quad (2)$$

ここで、j_i^{PBMT} は PBMT によって翻訳された i 番目の日本語文、j_i^{NMT} は NMT によって翻訳された i 番目の日本語文を意味する。ただし、単語一致率が 1、すなわち PBMT によって得られた日本語訳と NMT によって得られた日本語訳が表層で完全一致する場合、それらは言い換えではないので除外した。

4 言い換えアノテーション

表 1 に示すように、50 万文対の言い換え候補から単語一致率によって均等に 2,000 文対をサンプリングし、著者の 2 人がアノテーションを行った。アノテーションの基準を以下に示す。

[2] https://translate.google.co.jp/
[3] Google Sheets の GOOGLETRANSLATE 関数を使用
[4] https://dumps.wikimedia.org/enwiki/20160501/
[5] http://kheafield.com/code/kenlm/

表 1: 言い換え認識の評価用コーパスの統計

Jaccard	総文対数	標本数	正例	負例	誤訳	その他
[0.0, 0.1)	228	200	2	1	80	117
[0.1, 0.2)	2,117	200	11	14	147	28
[0.2, 0.3)	14,080	200	20	9	162	9
[0.3, 0.4)	51,316	200	24	15	161	0
[0.4, 0.5)	100,674	200	27	16	151	6
[0.5, 0.6)	134,101	200	34	16	142	8
[0.6, 0.7)	100,745	200	38	13	129	20
[0.7, 0.8)	55,610	200	53	12	131	4
[0.8, 0.9)	26,884	200	81	3	94	22
[0.9, 1.0)	8,071	200	73	3	56	68
[1.0, 1.0]	6,174	0	0	0	0	0
Total	500,000	2,000	363	102	1,253	282

正例 適切な日本語訳の対であり、言い換えである。
負例 適切な日本語訳の対だが、言い換えではない。
誤訳 少なくとも片方の文が不適切な日本語訳である。
その他 句読点の有無など些細な違いのみを含む文対や、ほとんどが固有名詞で構成されている文対。

まず著者の 1 人が上記の基準で 2,000 文対すべてのアノテーションを行った。そして、**正例**または**負例**のラベルが付与された文対に対して、別の著者の 1 人が再びアノテーションを行った。なお、アノテータ間の一致率（Cohen's kappa）は 0.60 と十分に高かった。少なくとも片方のアノテータが**誤訳**／**その他**とラベル付けした文対はコーパスに採用せず、アノテータ間で**正例**と**負例**のラベルが一致しなかった 89 文対については協議して最終的なラベルを決定した。その結果、363 文対の正例と 102 文対の負例からなる 465 文対の日本語の言い換え認識の評価用コーパスを構築した。

5 言い換えコーパスの分析

本研究で構築した言い換えコーパスの事例を表 2 のように分類した。まず、**文末表現**（特に**常体と敬体の変換**）が変化する正例が非常に多い。これは翻訳器の性質によるもので、本研究で使用した PBMT ツールは敬体を好み、NMT ツールは常体を好む傾向があった。次に多いのが**内容語の置換**による言い換えである。これを細分類したところ、「前例」と「先例」のような同義語の単純な置換（**語種の変化なし**）の事例が多く見られた。内容語の置換に含まれる**語種の変化あり**は、「規則」と「ルール」のように漢語が外来語に言い換えられているものである。また、**大きな単位での変換**の**フレーズ**とは、「宣戦布告する」と「戦争を宣言する」のように単語単位では言い換えでないにも関わらず、フレーズ単位で言い換えになっている事例である。最後に、**内容語の挿入・削除**は、「心配することは何もありません」と「心配することはありません」のように、自明な要素を挿入または削除する事例である。ここには「私は知らない」と「知りません」のような主語の省略も含まれる。

表 2: 獲得した事例の分類

分類	正例	負例	Total
内容語の置換	180	61	241
語種の変化なし	116	44	160
語種の変化あり	49	10	59
表記揺れ	14	5	19
片方向の含意関係	1	2	3
文末表現	143	34	177
常体と敬体の変換	122	16	138
アスペクト	12	7	19
ヴォイス	4	4	8
モダリティ	1	5	6
テンス	4	2	6
機能語の挿入・削除	54	8	62
機能語の置換	43	16	59
大きな単位での変換	22	21	43
フレーズ	20	19	39
文	2	2	4
内容語の挿入・削除	20	12	32
語順の変更	9	1	10
世界知識	2	5	7

表 3 に特徴的な事例を示す。本研究では単語一致率の低い非自明な正例を積極的に収集したため、#1 の大きな単位での変換が獲得できた。また、#2 の外来語を含む内容語の置換が多いことも、機械翻訳を用いて言い換え候補を収集した本研究の特徴である。#3 は流暢な誤訳の事例であり、「ジェネリック医薬品」という固有名詞を PBMT が普通名詞として翻訳している。#4 のように、どちらの機械翻訳も "Why do you work so hard?" の日本語訳として妥当かつ流暢であるが、言い換えではないという例も見られた。#5 は単語一致率の高い非自明な負例であるが、表 1 に示したように、このような事例は本手法では獲得が難しい。#6 は、#5 と同じく**片方向の含意関係**にある単語対を含むが、「連邦議会」がアメリカの議会を指すという**世界知識**および「オクラホマ州」がアメリカの州であるという世界知識を用いると、この文脈では「連邦議会」と「議会」の間に共参照の関係が成り立つことがわかり、文単位で言い換えになっていると判定できる。#7 は、言い換えではないフレーズ単位の変換の

表 3: 獲得した言い換えの事例

	Jaccard	ラベル	PBMT	NMT	分類
#1	0.07	正例	めったに使われることはありません。	まれに使用されます。	フレーズ
#2	0.60	正例	彼は共和党のメンバーでした。	彼は共和党の一員だった。	語種変化
#3	0.12	負例	これは、一般的な薬として利用可能です。	ジェネリック医薬品として入手できます。	世界知識
#4	0.15	負例	なぜあなたは一生懸命働くのですか？	どうしてそんなに頑張ってるの？	文
#5	0.80	負例	米国は 1819 年にスペインからフロリダを獲得しました。	米国は 1819 年にスペインからフロリダを買収した。	含意
#6	0.91	正例	彼女は 1921 年以来、オクラホマ州から連邦議会に選出された最初の女性です。	彼女は 1921 年以来オクラホマ州から議会に選出された最初の女性です。	含意･世界知識
#7	0.40	負例	「カンフーパンダ」は、批評家の称賛を受けました。	「カンフーパンダ」は批判的な評価を受けました。	フレーズ
#8	0.30	正例	1985 年に彼女は第一子を出産しました。	1985 年、彼女は最初の子供を産んだ。	内容語等

事例である。「批評」と「批判」は単語単位では言い換えであり、「評価」と「称賛」も上位下位関係にある意味的に近い単語対であるが、「批評家の称賛」と「批判的な評価」はフレーズ単位では同義ではない。#8 は、多くの言い換え関係の組み合わせの例である。機能語の置換、フレーズ単位の変換、内容語の置換、常体と敬体の変換の 4 つの変換が含まれている。

6 おわりに

本研究では、日本語の言い換え認識のための評価用コーパスを構築した。我々は、複数の機械翻訳を用いて言い換え候補の収集コストを抑え、単語一致率の低い非自明な言い換え事例を積極的に収集した。

今後の課題は、正例と負例のバランスの改善である。本研究で構築したコーパスは、正例が全体の 78.1%を占めており、非常に多い。そのため、全ての事例に対して「同義」と回答する単純な手法でも F 値 87.7 を達成できてしまう。また、本研究では単語一致率の低い正例を積極的に収集したが、一方で単語一致率の高い負例は収集できていない。そのため、単語一致率の高い事例に対しては、表層的な手掛かりのみで問題がある程度解ける [2] という先行研究の課題が解決できていない。単語一致率の低い事例に対するアノテーションを進めることでコーパス全体の正例と負例のバランスはある程度改善できると考えているが、本研究で注目しなかった単語一致率の高い非自明な負例の収集方法も今後の重要な課題である。

参考文献

[1] William B. Dolan and Chris Brockett. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proc. of IWP 2005*, pp. 9–16, 2005.

[2] 藤田篤, 柴田知秀, 松吉俊, 渡邉陽太郎, 梶原智之. 言い換え認識技術の評価に適した言い換えコーパスの構築指針. 言語処理学会第 21 回年次大会ワークショップ「自然言語処理におけるエラー分析」発表論文集, pp. 1–11, 2015.

[3] Regina Barzilay and Kathleen R. McKeown. Extracting Paraphrases from a Parallel Corpus. In *Proc. of ACL 2001*, pp. 50–57, 2001.

[4] Bo Pang, Kevin Knight, and Daniel Marcu. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proc. of NAACL 2003*, pp. 102–109, 2003.

[5] David Chen and William Dolan. Collecting Highly Parallel Data for Paraphrase Evaluation. In *Proc. of ACL 2011*, pp. 190–200, 2011.

[6] Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. Déjà Image-Captions: A Corpus of Expressive Descriptions in Repetition. In *Proc. of NAACL 2015*, pp. 504–514, 2015.

[7] Tomoyuki Kajiwara and Kazuhide Yamamoto. Evaluation Dataset and System for Japanese Lexical Simplification. In *Proc. of ACL-IJCNLP 2015 SRW*, pp. 35–40, 2015.

[8] Tomonori Kodaira, Tomoyuki Kajiwara, and Mamoru Komachi. Controlled and Balanced Dataset for Japanese Lexical Simplification. In *Proc. of ACL 2016 SRW*, pp. 1–7, 2016.

[9] Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Shuming Shi, Teruko Mitamura, Noriko Kando, Hideki Shima, and Kohichi Takeda. Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10. In *Proc. of NTCIR 2013*, pp. 385–404, 2013.

[10] Preslav Nakov, Francisco Guzman, and Stephan Vogel. Optimizing for Sentence-Level BLEU+1 Yields Short Translations. In *Proc. of COLING 2012*, pp. 1979–1994, 2012.

[11] Anna Kozlova, Mariya Shmatova, and Anton Frolov. YSDA Participation in the WMT'16 Quality Estimation Shared Task. In *Proc. of WMT 2016*, pp. 793–799, 2016.