テキスト正規化技術を用いたCGM日本語テキスト翻訳

笠原 要 斉藤 いつみ 浅野 久子 片山 太一 松尾 義博

NTT メディアインテリジェンス研究所
kasahara.kaname@lab.ntt.co.jp

1. はじめに

マイクロブログや口コミサイト等の消費者生成メディア(Consumer Generated Media, CGM)で出現する多様な表現のテキストを正規化することが、日本語の統計機械翻訳処理に有効か検討した結果を報告する。

投稿者の日記やつぶやき、新商品やサービスに関する口コミ情報等のCGMコンテンツに含まれるテキスト(以下、「CGMテキスト」と呼ぶ)には、新聞等マスメディアからは発信されない、消費者の生の声や地元情報が含まれ、商品購入時の参考情報となっている。多くは投稿者の母国語で記述されているため、外国の観光や製品に興味あるが母国語が異なる消費者には利用障壁が高い。テキスト規模が膨大なために、予めCGM事業者が翻訳して提供することも容易ではないため、機械翻訳の応用効果が高い分野と期待される。

CGMテキストは、新聞記事や特許文等の機械翻訳研究で主として扱われているテキストと較べて文表現の多様性が高い。新聞記事では不特定多数の読者が想定されるので、訓練を受けた記者が一般常識的な語彙や表記、語用で記事執筆・構成する。一方CGMでは、少数の身内や、読者すら前提としない投稿も多い。そのため、国語辞典に記載がない多様な単語表記や、身内にのみ通じる言い回しや語の省略等含まれる。また、twitter等のコミュニケーションツールでは、顔文字のような、文字を使ったノンバーバルな表現がテキスト中で頻出する。

CGMテキストを統計機械翻訳処理する場合には、このような表現を含む対訳コーパスが必要となる。twitter等のマイクロブログから対訳データを自動構築する方法[1-2]やクラウドソーシングで収集する方法[3]等提案されているが、日本語については、頑強な形態素解析器検討のためのアノテーション方法の検討段階[4]であり、日本語のCGMテキスト翻訳は容易には行えない。そこで、既存の対訳データを活用し、入力のCGMテキストの言語表現を揃える"正規化"を行い、翻訳処理する方法を検討する。翻訳結果が変化しないように正規化できるならば、多様な表現に起因するデータ・スパースネスの問題が軽減される。さらに、統計機械翻訳でのモデル学習やデコード処理前の解析誤りも減少し、訳文の品質向上も期待される。CGMテキストの正規化としては、様々な処理段階があり、それに対する種々の方式が提案されている。例えば、単語表記の正規化では、崩れた日本語テキストの形態素解析、表記正規化手法が提案されている(例えば、[5-7])。また、述語を構成する単語列を正規化して簡易な表現に変換する方法[8]や、それを形態素解析辞書に登録して翻訳を行う方法[9]が提案されている。これ以外にも、CGMテキストで散見される、語の省略や顔文字等について対訳コーパスでの文表現に近づくように変換する必要が考えられるが、これらの"正規化"処理の何れを統計機械翻処理とどのように結合させると有効であるかは明らかではない。

そこで本稿では、統計機械翻訳を行う際のCGMテキストの特徴を整理する。これらを手動で正規化した場合の翻訳精度の向上の度合いを確認し、自動化した時の有効性について述べる。

2. CGMテキストの正規化

本章では、CGMテキストにおける多様な表現の特徴及び、これを正規化することの有効性を説明する。

2.1. 単語の表現

意味を同じくし音が同一、あるいは類似した形態素について様々な表記が用いられる。以下、日本語twitterにおける例を挙げて説明する。

(日本語tweet 例1)

半分の量でいい**なぁ**。/それだけ旨いってことなんだろう**な**。/こりゃまた綺麗なゴールだ**なあ**。センタリングもいい**なー**

同一の終助詞「な」について、類似した音の表記揺れとして出現している例である。同一の音でも平仮名/片仮名/漢字の使われ方が一定でない場合もある(例:「ヤバ/イ」、「ヤバ/い」、「やば/い」)。さらには、あて字や誤字も大きな意味での形態素の表現の多様性と捉えられる。

形態素解析処理で「なー」が形態素「な」と「ー」に分割されたり、隣接する形態素と一体になってしまうと、統計機械翻訳処理で適切な統計情報を正しく参照できなくなり、翻訳品質が低下する。また、「なー」と正しく分かち書きされても新聞記事のような一般常識的なコーパスでは出現頻度が低いと思われ、データ・スパースネスが解消できないので期待する翻訳結果が得られない。そのため、崩れた形態素表記を正しく分かち書きすると共に表記を正書化(上例では「なあ」→「な」)できる形態素解析処理(例えば文献、[5-7])を用いる必要

がある。一方、正書表記に変換する際に、音が同じ、あるいは類似した異義語に誤変換(例えば、助詞「なあ」を人名「菜愛」に変換)されると、翻訳精度が低下する恐れがある。

2.2. 顔文字の表現

CGM テキストでは、投稿者や読者の感情を表現するために、表情を文字列で表現する顔文字が多用される。

(日本語 tweet 例 2)

> ルールが守れないなら帰れよ!!ヽ(o`Д´o)ノ
> でも、当たったら何かご馳走して～(笑)
> イライラしてたらケーキ買ってきてくれた*・゜゜・*:.。..。.:*・'(*゜▽゜*)'・*:.。..。.:*・゜゜・*

上記の「ヽ(o`Д´o)ノ」等が顔文字である。「(笑)」は人の表情を合していないが、文中の他の語と統語的な結びつきは無いために本稿では同列と扱う。顔文字は、通常の形態素と同種の文字列が用いられるので、形態素解析処理で複数の形態素として出力されてしまうと、続く係り受け処理や翻訳処理の精度低下に結びつく。そのため、顔文字が辞書に登録されている形態素解析器の利用や顔文字の抽出を行うことが必要となる。

2.3. 述部の表現

CGM テキストは口語的表現が多く、述部を複数の動詞、助動詞、助詞で表現する事例が散見される。

(日本語 tweet 例 3)

> タイトスカートと合わせるとちょうど**よくなるような**･･･／最初に返すって言葉を言ったわりにいざとなったら**逃げるわけだから**。／5 人のデビューからの成長記録からテストが**出ればいいのにね**。／彼氏がいるからこの部活入る、ってふざけてんの？って**思うんだけど**。

新聞記事等では文は簡潔に記述されるので、対訳コーパスで上記のような単語 n-gram の出現頻度が低く、そのまま翻訳処理を行っても、期待する訳文が得られない恐れがある。

これに対して、文献[8]のような述部の単語列を意味分類した辞書を用いて、簡潔な述部に変換(例えば、「思うんだけど」を「思う」に変換)する技術を用いれば、対訳コーパス中の述部と一致する可能性が高まる。一方、述部の正規化では、「けど」のような訳文に影響がある(例えば、'but'が訳文に加わる)形態素も省かれる恐れがある。

2.4. 文の表現

CGM テキストでは、新聞と異なり各投稿者の心情を制限しないため、文末で体言止めや「…」が用いられる。また、会話文に近い表現も多いので、主語や助詞の省略も見かけられる。

(日本語 tweet 例 4、括弧内は補完結果)

> 観客席**(は)**ガラガラ**(だ)**…／気分**(が)**悪いね！ヽ(`Д´)ノ／**(私は)**さっきみたいなクロス**(が)**好き／午前はベビーの耳鼻科**(だ)**。

これら形態素の省略を補完しないで処理すると、係り受け解析誤りが発生する。統計機械翻訳で統語解析が用いられる場合は、翻訳誤りに繋がる場合がある。省略補完処理を行うことにより上記問題の解消が期待される。一方、補完誤りは解析誤りにつながるため、翻訳精度が低下する恐れがある。

2.5. 統計機械翻訳のための正規化処理

上記で述べたことを整理した正規化処理方法を表1に挙げる。それぞれが翻訳処理での解析誤りの低減に効果があると思われる。また、顔文字以外の正規化では、学習データのデータ・スパースネスの問題が軽減されると期待される。

各正規化は、個別に行うことが可能であるが、ボトムアップ的に続けて行う(例えば、(顔文字、単語、述部、文)ことが効果的と推測される。

表1 CGM テキスト翻訳のための正規化処理

正規化		頻度	解析	副作用
単語	対訳データに出現する表記に統一	○	○	同音異義語
顔文字	文から除去	－	○	－
述部	簡潔に表現	○	○	述部意味変化
文	省略された語を補完	○	○	補完誤り

3. 実験

日本語の CGM テキストの入力文に対し、前章で述べた正規化の何れが有効であるか実験的に調べた。まず、人手でいくつかの正規化処理を行った場合の翻訳処理での適用性について予備実験で調べた。次にいくつかの自動正規化手法を適用して評価し、その効果を確認した。

3.1. 人手での正規化

日本語 tweet 465 文と英語の対訳文を用いて予備実験を行った。まず、各日本語文について、人手で下記の正規化処理行った(表2)。

表2 人手正規化処理手順

① **単語**音が同一/類似し意味が同じ単語表記の内で、常識的に用いられると考えられる表記に変換(方言を含む)
② **述部** 述語と関わる単語列について、同一の意味を表す簡潔な表現に変換
③ **文** 主語や助詞等が省略されていると推測出来る場合は、文のみの情報から省略された形態素を補完
④ **顔文字** 文中に出現する顔文字を構成する文字列を除去

それぞれの正規化は、前の正規化結果の文に対して行った。②の述部正規化手法としては、「反発をくらう」ような名詞、助詞、動詞で表現される構文のパタンを分類し同一の意味である簡潔な表現(例えば「反発される」)に変換する方式[1]を用い、その出力結果を作業者が訂正することで行った。また、①から③までは1名の作業者で行い、③の最後に別の作業者が不足/不適切な処理を修正した。

CGM テキストに対して順に正規化を行った例を表3に示す。例1では顔文字は含まれていないが、それ以外の正規化が各々行われている。例2では、主語の補完及び文末の顔文字が削除されている。

表3 日本語 tweet 文の人手正規化例

例	操作	操作結果
1	(原文)	つくえほしいねえ
	①表記正規	机欲しいね
	②述部正規	机欲しい
	③省略	私は、机が欲しい
2	原文	ステーキ丼とかずいぶん豪華な物を久しぶりに食べた_(ˆ)_
	③省略	私は、ステーキ丼とかずいぶん豪華な物を久しぶりに食べた_(ˆ)_
	④顔文字	私は、ステーキ丼とかずいぶん豪華な物を久しぶりに食べた

得られた各段階の正規化テキストについて、統計翻訳で対訳文を生成した。対訳モデル、言語モデル作成に用いた対訳データは、新聞記事、辞書の用例等で構成される 100 万の対訳文対である。統計機械翻訳プログラムは、英中韓から日本語への翻訳システム[10]をベースに 事前並べ替えに基づく日英翻訳方式[11]を行えるよう改造したものを用い、チューニングは MERT[12]で行った。翻訳結果の評価は、日本語 tweet 原文に対して予め作成しておいた英語対訳文と比較し、BLEU-4[13]、RIBES[14]のスコアを算出した。

実験結果を表4に示す。処理ステップ毎にはスコアの向上が見られ、ここに取り上げられた手続が機械翻訳の事前処理として有効であることが示唆される。述部正規化処理ではスコアが低下している。文献[8]の正規化は、日本語での情報抽出を志向して最適化されており、機械翻訳での使用について最適化されていないため、翻訳で有効な述部の分類を調整するなどの検討が今後の課題である。

また、評価値全体が新聞記事や特許文等の統計機械翻訳と較べて BLEU スコアで1桁程度小さくなっており、その原因の解析も課題である。

表4 人手正規化の効果

正規化処理	BLEU	RIBES
(Tweet 原文)	4.81	0.491
① 単語	5.25	0.499
② 述部	4.56	0.485
③ 文	5.55	0.549
④ 顔文字	5.85	0.550

3.2. 自動正規化

予備実験で明らかとなった翻訳への適用に評価があった正規化処理の内で、自動化した場合にも翻訳精度向上するか、実装可能であった単語表記正規化と顔文字処理について検証した。

○ 単語正規化

単語表記の正規化処理として、文献[7]の表記正規化と形態素解析を同時に行う方式を用いた。日本語 twitter での崩れ表記に対する正規表記のペアデータを学習データとして文字列の正規化パタンを獲得して形態素解析ラティスを拡張する点が特徴である。実験では JTAG[15]辞書の標準表記を使用した。

適用において、テストデータ、対訳データ共に正規化することで同一の語を同じ表記に対応づけることができる。実験では、テストデータのみ正規化する簡易な方法として、対訳データの日本語文中で頻出する形態素の表記を正書表記として変換した。

○ 顔文字処理

顔文字記号は、文中の統語解析の対象とならないと考え、日本語文中から見出された顔文字については、一旦除外して翻訳処理を行った後に、訳文に追加することを行った。日本語形態素解析システム JTAG では、顔文字として品詞付けすることが可能である。顔文字の対訳辞書を構築して変換することが好ましいが、パタンとしての顔文字は固有の言語を超えて理解可能な場合が多いと考え、本稿では、日本語文で現れる顔文字を訳文にそのまま配置することとした。

3.3. 実験結果

学習モデル構築に用いる対訳データ及び翻訳方式は、正規化の人手評価と同一のものを用いた。テスト文

としては、日本語 twitter 文(N=2,000)に加え、レストランロコミサイトの文(N=1,000)から対訳文を作成して用いた。

適応検討の第一歩として、表記正規化、顔文字除去を個別、組合せて行った結果を表 4 に示す。単語表記の正規化ではロコミ文のテストデータでは BLEU スコアが向上している一方、それ以外では、スコアが低下している。テスト文の分かち書きは正しいが、選択された表記が対訳データで出現する単語表記と一致しない場合がある等の理由と推測される。顔文字については、tweet 文で顕著にスコア向上している。不特定多数を対象として投稿されたロコミ文と比べて少数の知り合いを意識して投稿される tweet 文では、顔文字の出現頻度が多いためと推測される。2つの正規化を組み合わせた結果については、各々のスコアを上回っていないため、効果的に組み合わせる方法の検討も課題である。また、評価値全体は、人手評価の時と同様に低い。

表 4 正規化処理の適用結果

正規化処理	Tweet (N=2,000)		ロコミ(N=1000)	
	BLEU	RIBES	BLEU	RIBES
(原文)	4.24	0.438	5.49	0.547
①単語表記	3.74	0.416	**5.59**	0.539
②顔文字	**9.48**	**0.487**	5.50	0.547
①+②	9.08	0.468	5.57	0.539

4. おわりに

本稿では、日本語CGMに含まれるテキストを統計機械翻訳処理する際に、単語表記や文表現を正規化することが有効であるかについて検討を行った。日本語 tweet 文に対して単語、述部、文、顔文字の正規化を段階的に人手で行った結果、翻訳精度の向上につながることを確認した。そして、単語表記正規化及び顔文字除去の自動処理を適用した結果、翻訳精度の一部向上が見られることを確認した。それぞれの正規化を統計機械翻訳に適用するための調整は課題である。述部正規化及び文正規化の有効性については、今回検証できなかったので、これについても検証し、他の正規化方法と組合せて最適な CGM テキストの翻訳方法を見出す検討を行う予定である。

また、原文及び正規化したテスト文の翻訳結果のスコアが、研究が進んでいる新聞記事・特許翻訳と較べて相当低くなっている。今回はテスト文それぞれに対して1つの対訳文を作成して検討したが、CGM テキストではゼロ代名詞等の形態素の省略が多いため、意味が異なるが適切な複数の訳文が考えられる場合も多いためと推測される。翻訳文の人手評価なども加え、正規化の効果を総合的に評価加える事も課題である。

参考文献

[1] L. Jehl, F. Hieber, and S. Riezler, "Twitter translation using translation-based cross-lingual retrieval,"The Seventh Workshop on Statistical Machine Translation, 2012, pp. 410-421.

[2] W. Ling, G. Xiang, C. Dyer, A. Black, and I. Trancoso, "Microblogs as Parallel Corpora," ACL-2013, 2013.

[3] W. Ling, L. Marujo, C. Dyer, A. Black, and I. Trancoso, "Crowdsourcing High-Quality Parallel Data Extraction from Twitter." The Ninth Workshop on Statistical Machine Translation (WMT).

[4] 鍛治, 吉永,喜連川, "マイクロブログに対する形態素・正規化情報のアノテーション," 言語処理学会年次大会, 2014, pp. 908-911.

[5] 工藤, 市川, D. Talbot, 賀沢, "機械学習に基づくマイクロブログ上のテキストの正規化," 言語処理学会年次大会,2012, pp. 1276-1279.

[6] 佐々木, 水野, 岡崎, 乾, "機械学習に基づくマイクロブログ上のテキストの正規化," 人工知能学会全国大会, 2013.

[7] I. Saito, K. Sadamitsu, H. Asano, and Y. Matsuo, "Morphological Analysis for Japanese noisy text based on character-level and word-level normalization," COLING 2014, 2014.

[8] T. Izumi, K. Imamura, G. Kikui, and S. Sato, "Standardizing complex functional expressions in Japanese predicates: Applying theoretically-based paraphrasing rules," The Multiword Expressions: From Theory to Applications, 2010, pp. 64-72.

[9] 坂本, 園尾, 田中, 釜谷, "機能語相当の日本語 MWE の段階的まとめあげによる 統計的機械翻訳の精度向上," 言語処理学会 第 19 回年次大会 発表論文集, 2013, pp. 30-33.

[10] 須藤, 鈴木, 塚田, 永田, 星野, 宮尾, "語順の入れ替えに着目した特許の統計翻訳: 事前・事後並べ替えによる高精度な英日・日英翻訳 (機械翻訳技術の向上)," Japio year book, 2013, pp. 292-296.

[11] 星野, 宮尾, 須藤, and 永田, "日英統計的機械翻訳のための述語項構造に基づく事前並べ替え," 言語処理学会年次大会, 2013.

[12] F. J. Och, "Minimum error rate training in statistical machine translation," The 41st Annual Meeting on Association for Computational Linguistics, 2003, pp. 160-167.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," The 40th annual meeting on association for computational linguistics, 2002, pp. 311-318.

[14] 平尾, 磯崎, 須藤, K. Duh, 塚田, 永田, "語順の相関に基づく機械翻訳の自動評価法," 自然言語処理, vol. 21, no. 3, pp. 421-444, 2014.

[15] T. Fuchi and S. Takagi, "Japanese morphological analyzer using word co-occurrence: JTAG," COLING, 1998, pp. 409-413.