表記ゆれの統計的機械翻訳への影響

髙橋 寛治　　山本 和英
長岡技術科学大学
{takahashi, yamamoto}@jnlp.org

1　表記ゆれの影響を明らかにしたい

日本語に表記ゆれは多く、Web 文書中の 10%の語において表記ゆれが見られる [8][9]。表記ゆれの解消は語彙数を減らし単語の組み合わせ数を減少させるため、後段処理の性能向上が期待できる [5]。我々は表記ゆれの解消を目指して解析器を構築する [7] とともに、表記ゆれの影響を調査している。

表記ゆれの影響の調査対象には統計的機械翻訳を用いる。被覆率向上を目指した換言処理 [1] やマイクロブログに見られる崩れ語の正規化 [2] が統計的機械翻訳の性能を向上させるため、表記ゆれの解消も性能を向上させると考えた。我々 [3] は以前に PBMT(句に基づく統計的機械翻訳) への表記ゆれの影響を調査したが、いくつか問題がある。活用を考慮していない上に、異なる単語単位での評価 [1] を行っている。

我々は、上記問題を改善して実験を行ったが性能は改善されなかった [4]。この理由に、表記ゆれの数を調査したところコーパス全域に影響を及ぼすほどの表記ゆれが存在しないためと考えた。それでは表記ゆれが多数含まれるコーパスの場合は、機械翻訳の性能が向上するのだろうか。また、入力に表記ゆれが含まれた場合性能は低下しないのであろうか。

本研究では、以下の 3 つの調査を行うことで、表記ゆれの統計的機械翻訳への影響を明らかにする。

- PBMT において、翻訳モデルの学習と言語モデルの学習に異なるコーパスを用いた際の影響 (3 章)
- 低頻度な表記が PBMT および NMT(ニューラル機械翻訳) の出力に与える影響 (4 章)
- 表記ゆれを多数含むようにコーパスを加工した際の機械翻訳への影響 (5 章)

[1]BLEU や RIBES は単語単位に依存した評価尺度であるため、単語単位が短い場合は BLEU を多く計上する傾向がある

2　表記ゆれ

2.1　対象となる表記ゆれ

日本語解析システム雪だるま (1) を用いて表記ゆれを解消する。雪だるまの表記ゆれ解消の対象 [2] は普通の日本語とし、崩れ語や固有表現などは考慮していない [7]。対象となる表記ゆれを例とともに以下に示す。

字種 りんご, リンゴ, 林檎

外来語 コンピュータ, コンピューター

漢字間の異なり 附属, 付属

送りがなの異なり 受付ける, 受け付ける

部分的な漢字仮名交じり単語 改ざんする, 改竄する

その他 とても, とっても

2.2　雪だるまによる表記ゆれの解消

雪だるまによる表記ゆれ解消の概略を説明する。雪だるまは形態素解析辞書のエントリーを ID 管理することで表記ゆれを解消する。MeCab-UniDic が出力した形態素 ID 列に対して辞書を用いて表記ゆれを解消する。具体的には、「りんご」「リンゴ」「林檎」が同じ単語 ID を持っているということである。

ひらがなに対する安易な表記ゆれ解消は誤る可能性がある。例えば、「月光 (がっこう)」のような記述がある際に、「がっこう」は「学校」として利用することが一般的であると考え、辞書に表記ゆれとして登録されている。これを表記ゆれ解消すると「月光 (学校)」となる。他には、形態素解析誤りに起因する問題として分割誤りによる誤った表記ゆれ解消がある。例えば、「てんじん」が「てんじ　ん」と誤分割され、表記ゆれ辞書の都合上「展示　ん」と解析される。表記ゆれ解消には誤りがあることを念頭に置きつつ実験および考察する。

[2]http://snowman.jnlp.org/snowman/target に詳細を記載。

3　言語モデルと翻訳モデルを別コーパスで学習した影響

3.1　目的

表記ゆれが多く含まれるコーパスにおいて、表記ゆれ解消の影響を調査することを目的とする。そのために、翻訳モデル学習用コーパスと言語モデル学習用コーパスに別の種類のコーパスを用いて影響を調査する。

単一種類のコーパス内で表記ゆれを解消しても機械翻訳の性能は向上しなかった [4]。これは実験に用いるコーパス内に含まれる表記ゆれが少ないためであった。実験では単一のコーパスを用い評価を行うが、現実には多種多様の文書を投入する。これにより表記ゆれが多くなり、表記ゆれ解消の影響が大きくなると考えられる。

3.2　実験設定

翻訳モデルと言語モデルの学習に別のコーパスを用いることのできる PBMT を用いて比較を行い、影響を調査する。ここで表記ゆれは日本語を対象としているため、日本語言語モデルを用いる英日翻訳で検証を行う。

言語モデルは、現代書き言葉均衡コーパス (BCCWJ)(2) すべてを KenLM を用いて 5-gram モデルで学習する。翻訳モデルの学習、テストセットおよびチューニングには Wikipedia 京都関連テキスト (KFTT)(4) を用いる。PBMT には Moses(3) を用いる。チューニングなどのパラメータはデフォルト値を利用する。評価は BLEU および RIBES を用いる。なおチューニングは 5 回行い、それぞれに翻訳を行う。評価は 5 つの訳出に対して行い、平均をとる [3]。

3.3　結果と考察

表 1 に結果を示す。参照訳の表記ゆれが解消されているため、必ずしも正確な比較ではないことに注意されたい。数詞を除く表記ゆれ解消が最も低いスコアとなり、数詞を含めた表記ゆれ解消が最も高いスコアとなった。

表記ゆれ解消（数詞を除く）はベースラインに比べて両評価尺度のスコアが低い。これは表記ゆれ解消の誤りが寄与していると考えられる。一般にひらがなで記述されたものを形態素解析すると、解析誤りが起きやすい。KFTT には読みがなを記述するものが多く、これらが解析誤りとなりスコアが低下したと考えられる。対処としては、ひらがな語を表記ゆれ解消の対象から除くことや、ひらがなの表記決定を語義曖昧性解消と捉えたかな漢字換言システム [6] で処理することで改善が期待できる。

表 1: 学習コーパスの表記の差異が与える影響

実験設定	BLEU	RIBES
ベースライン	16.8	64.6
表記ゆれ解消（数詞）	18.4	66.4
表記ゆれ解消（数詞を除く）	16.6	64.5
表記ゆれ解消	18.5	65.6

数詞を含めた表記ゆれ解消は、性能向上に寄与している。両コーパスに含まれる数詞のユニグラム頻度を調査したところ、KFTT では 6.7%、BCCWJ では 3.8%を占めている。頻出する単語を統制した場合、性能は向上することが明らかとなった。

結果および考察から、処理対象のコーパスに頻出する表記ゆれの解消を行うことで性能が向上すると言える。また、複数コーパスを利用する場合は表記ゆれの解消を行うことで性能が向上する。

4　低頻度の表記ゆれの影響

4.1　目的

あるコーパスで学習した翻訳システムに対して、入力文が表記ゆれを起こすことでどのような影響があるかを調査する。統計的機械翻訳は単語や単語列の頻度を元に翻訳を行う。ここで同じ単語だが下記のように表記が異なる語は別の語として扱われる。

例 1) ほとんど、殆ど

例 2) たなばた、棚機

表記ゆれが影響しないならば入力文が表記ゆれを起こしても評価は変わらず、影響する場合は評価値が変化するはずである。これを明らかにするために、表記を変更した入力の影響を調査する。

[3] 事前実験ではチューニングごとに BLEU 値が 3 ポイント前後変化した。そのため平均をとった。

4.2 実験設定

入力文中の語の表記を変更する。入力文中の語のうち、表記ゆれを学習コーパス内で持つ語を学習コーパス内で低頻度な表記へと置換する [4]。例を示すと、「名付け:9」「名付:4」（数値は頻度を表す）という表記ゆれがが学習コーパスに存在する場合、入力文中の「名付け」を「名付」に変更する。

低頻度な表記へ置換したものを入力文として、翻訳結果を BLEU および RIBES で評価する。コーパスには KFTT を用いる。NMT および PBMT で比較する。NMT には注意型 Sequence-to-Sequence モデルが実装されたもの (5) を、PBMT には Moses(3) を利用する。両方の機械翻訳システムのパラメータはデフォルト設定値を利用する。

4.3 結果と考察

表 2 に結果を示す。NMT および PBMT の両方で、BLEU および RIBES によるスコアが約 1 ポイント低下している。

表 2: 低頻度な表記ゆれを入力させた日英翻訳

実験設定	NMT		PBMT	
	BLEU	RIBES	BLEU	RIBES
ベースライン	21.6	70.1	18.6	66.5
低頻度な表記	20.6	69.0	16.9	65.0

単語の表記が低頻度な表記（単語）へ変更されると機械翻訳の性能は低下した。低頻度な表記は機械翻訳に悪影響を及ぼすことが分かる。低頻度語は限られた文脈で出現し、なおかつその頻度も低いため良い翻訳が確率的に導き出せないと考えられる。よってこのような結果が得られたと考えられる。

実際の機械翻訳システムでは様々な表記の文が入力される予期される。本実験の結果から、実システムにおける表記ゆれの解消は機械翻訳システムの性能低下を防ぐことができる。

[4] コーパス中に存在する表記に置換しないと、未知語扱いになり未知語の問題が発生する。

5 表記ゆれコーパスを用いた影響調査

5.1 目的

表記ゆれが多数含まれるコーパスを作成し、その影響を調査する。表記ゆれが多数含まれる場合は、機械翻訳の性能が低下するはずである。先行研究 [4] でも同様の検証を行ったが、すべての品詞の単語から無作為に表記を選択していたため、誤った表記が多数含まれていた。本稿では、誤った表記を可能な限り含まないコーパスで検証を行うことで、影響を調査する。

5.2 実験設定

実験に用いるコーパスの日本語側すべて（入力文、学習コーパス、チューニング文）に対して、表記ゆれを起こす処理を行う。具体的には、下記例 3 が入力されると表記ゆれを起こした文例 4 が出力される。

例 3) このリンゴはとてもおいしそうである。

例 4) この林檎はとっても美味しそうである。

表記ゆれを起こした文は、表記ゆれ辞書から無作為に表記を選択することで変換する。ここで、表記ゆれを起こす対象は、名詞と副詞 [5] とする。これは先にも述べたように、動詞や形容詞も多数表記ゆれが存在するが、活用がある語の表記の復元が十分な性能に現時点で達していないためである。名詞と副詞は表記ゆれをうまく生成できるため、処理対象とする。

評価は日英翻訳により行う。翻訳システムには 4 章と同じ NMT および PBMT を用いる。実験に利用するすべてのコーパス（学習コーパス、チューニングコーパス、テストコーパス）に対して、表記ゆれ処理を加える。

5.3 結果と考察

結果を表 3 に示す。NMT では約 0.5 ポイント、PBMT では約 1.4 ポイントのスコアの低下が見られる。

無作為に表記ゆれを起こした場合に性能が低下した。無作為に表記ゆれを起こすということは、機械翻訳が取り扱う単語数が増加するということである。また無作為であるため、増えた表記も一定数の頻度で現れる。

[5] 雪だるま品詞体系については、http://snowman.jnlp.org/snowman/part-of-speech を参照。

表 3: 無作為に表記ゆれを起こした日英翻訳

実験設定	NMT		PBMT	
	BLEU	RIBES	BLEU	RIBES
ベースライン	21.6	70.1	18.6	66.5
表記ゆれ	21.2	69.6	17.2	65.0

これにより、表記ゆれが機械翻訳の性能を低下させたと考えられる。

表記ゆれの解消が翻訳性能に影響しなかったのは、実験コーパス中に含まれる表記ゆれが少ないからである。表記ゆれが大量に含まれるコーパス（無作為に表記ゆれを起こした場合）は、機械翻訳の性能を低下させた。このことから、表記ゆれをたくさん含むコーパスを扱う際には、表記ゆれを解消することで性能が向上がすると言える。

6 まとめ

本稿では表記ゆれの解消が統計的機械翻訳へ与える影響について３つの観点から調査した。性能の向上は見られなかったが、表記ゆれが与える影響について明らかにした。表記ゆれの統計的機械翻訳への影響に対する結論は以下の 3 点である。

- 複数コーパスの利用など現実の機械翻訳を考えた場合、表記ゆれを吸収したほうがいい
- コーパスに含まれる表記ゆれが少ない場合は、表記ゆれ解消の影響は皆無に等しい
- 表記ゆれが多数含まれる場合は、表記ゆれを解消すべきである

表記ゆれは表出しにくい問題であるが、解析や処理誤りを引き起こす要因の一つである。今後も継続した表記ゆれの解消に取り組むことで、地道な改善を目指す。

謝辞

本研究は、平成 27〜31 年科学研究費補助金基盤 (B) 課題番号 15H03216 の助成を受けています。

使用したツールと言語資源

(1) 日本語解析システム「雪だるま」, (Yamamoto et al. 2015), 長岡技術科学大学 自然言語処理研究室, `http://snowman.jnlp.org/`

(2) 現代日本語書き言葉均衡コーパス (BCCWJ), Ver.1.1, 国立国語研究所.

(3) Open source statistical machine translation system Moses. `http://www.statmt.org/moses`

(4) The Kyoto Free Translation Task, Graham Neubig, `http://www.phontron.com/kftt`

(5) Sequence-to-Sequence Learning with Attentional Neural Networks, Yoon Kim, `https://github.com/harvardnlp/seq2seq-attn`

参考文献

[1] Chris Callison-Burch, Miles Osborne, Philipp Koehn, and Miles Osborne. Improved Statistical Machine Translation Using Paraphrases. *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 17–24, 2006.

[2] Pidong Wang and Hwee Tou Ng. A Beam-Search Decoder for Normalization of Social Media Text with Application to Machine Translation. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, No. June, pp. 471–481, 2013.

[3] Kazuhide Yamamoto, Yuki Miyanishi, Kanji Takahashi, Yoshiki Inomata, Yuki Mikami, and Yuta Sudo. What We Need is Word, Not Morpheme; Constructing Word Analyzer for Japanese. *Proceedings of the International Conference on Asian Language Processing*, pp. 49–52, 2015.

[4] Kazuhide Yamamoto and Kanji Takahashi. Japanese orthographical normalization does not work for statistical machine translation. *Proceedings of the International Conference on Asian Language Processing*, pp. 133–136, 2016.

[5] 山本和英. 日本語の表記ゆれ問題に関する考察と対処. *JAPIO YEAR BOOK 2015*, pp. 202–205, 2015.

[6] 山本和英, 三上侑城. 語義曖昧性解消としてのかな漢字換言システムの開発. 言語処理学会第 22 回年次大会, pp. 180–183, 2016.

[7] 山本和英, 高橋寛治, 椙澤優希, 西山浩気. 日本語解析システム「雪だるま」第 2 報〜進捗報告と活用形態素の導入〜. 電子情報通信学会テキストマイニングシンポジウム, 信学技報, pp. 63–68, 2016.

[8] 佐藤理史. 異表記同語認定のための辞書編纂 (解析). 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2004, No. 47, pp. 97–104, 2004.

[9] 小椋秀樹. コーパスに基づく現代語表記のゆれの調査 — BCCWJ コアデータを資料として. 第 1 回 コーパス日本語学ワークショップ, pp. 321–328, 2009.