日本語の表記ゆれ問題に関する考察と対処

On orthographical variants problem and our solution

長岡技術科学大学准教授 **山本　和英**

PROFILE 豊橋技術科学大学大学院工学研究科博士後期課程システム情報工学専攻修了。博士（工学）。1996 年～ 2005 年（株）国際電気通信基礎技術研究所（ATR）、2002 年～現在まで長岡技術科学大学、現在准教授。自然言語処理、及び日本語教育（ツール作成）の研究に従事。

yamamoto@jnlp.org

1 日本語における表記ゆれ

自然言語が持つ特徴の一つに、表現の多様性がある。同一の事象や概念を表現するために、自然言語では様々な表現が可能である。我々人間はそれを意図的に使い分けたり、また逆に意識せずに任意の一表現を使ったりする。

言語表現の多様性は、諸外国語よりも日本語においてより顕著である。本稿では他言語との比較が目的ではないので詳細は省略するが、少なくとも日本語では表現の多様性だけでなく、下記に示すような表記の多様性、いわゆる「表記ゆれ」が数多く存在するからである。

- 文字種の多様性：
  「りんご」（ひらがな）
  「リンゴ」（カタカナ）
  「林檎」（漢字）
- （漢字における）字体の異なり：
  「付属」
  「附属」（旧字体）
- 送り仮名の異なり：
  「受け付ける」
  「受付ける」
- 外来語表記の異なり：
  「コンピューター」
  「コンピュータ」
- 略語：
  「取扱説明書」
  「取説」

以上のような表記ゆれは、自然言語処理を難しくしている。表記ゆれが増すことで処理すべき語彙が不必要に増大し、これが自然言語処理の様々な問題をより難しくしている。またこれは、産業界におけるテキスト再利用（産業日本語）という観点からも重大な問題である。表記ゆれが適切に吸収、あるいは同一性の認識がなされていないと、文書の検索などに大きな支障を及ぼすことが予想される。

本稿では、以上のような日本語の表記ゆれの問題について議論する。まず次章において表記ゆれの影響の大きさについて考察し、これに対する従来の取り組みと我々の取り組みについて紹介する。

2 表記ゆれの影響

まず、日本語にはどの程度表記ゆれがあるのか。これについては、小椋が現代日本語書き言葉均衡コーパス（BCCWJ）を対象に調査を行っている [1]。これによると、BCCWJ コーパスに対して表記ゆれの割合を調査した結果、書籍で 10.8%、Web で 10.3%、雑誌で 9.0% の形態素が表記ゆれであった[1]。自然言語処理でよく対象となる Web テキストにおいておよそ 1 割の単語が何らかの表記ゆれを持つというのは大きな意味がある。

ここで簡単な試算をする。仮に入力テキストの 1 割に

1　ただし、文献 [1] で言う表記ゆれは、「上げる－挙げる－揚げる」などの異字同訓も含んでおり、本研究における表記ゆれよりは対象が広いことに注意する必要がある。

表記ゆれがあると自然言語処理でどのような影響があるのだろうか。例えば自然言語処理では単語の共起情報を使うことがあるが、もし単語の１割が表記ゆれであれば、単純計算で全共起対の約２割は何らかの意味で表記ゆれの影響を受けることになる。述語と格要素の関係を見る格解析・述語項構造解析も、格関係に表記ゆれの影響がほぼないと見做すことができるので、表記ゆれに影響するのは述語と格要素の二項であり、影響の大きさは共起関係と同等である。

n-gram 統計を使う場合はさらに影響は大きくなる。例えば形態素 3-gram を用いるのであれば $0.9^3 \fallingdotseq 0.73$、さらに統計的機械翻訳などの際に見受けられる形態素 5-gram の場合は $0.9^5 \fallingdotseq 0.59$ となって約半数弱の n-gram 統計に不要な影響を与えることになる。

以上の試算より、表記ゆれ問題は決して小さな問題ではない。概算ながら上記で試算した割合は全く看過できない割合であり、手法・モデルの改善によって得られる効果よりもはるかに大きな効果が得られる可能性がある。

3 従来の取り組み

以上のように、表現多様性に起因する表記ゆれ問題は決して小さな問題ではない。さらに、本問題は形態素レベルの問題であることから自然言語処理のほぼすべてのタスクに影響を与える。

これに対し、従来の形態素解析器がどのように対処してきたかについて概観し、問題がある場合は指摘する。

3.1 NAIST-jdic

NAIST-jdic は、形態素解析用辞書 IPADIC の ICOT 条項をクリアすると共に表記ゆれ情報、複合語情報を付与した形態素解析用辞書である。この辞書内には Diff_notation という項目があり、これによって表記ゆれのまとめ上げを行っている。しかし、我々の調査では、下記のように必要以上にまとめ上げている傾向が見られる。

- 空ける／明ける
- 感づく／勘づく
- 炒る／煎る／いる
- 甘い／美味い／あまい
- 帰る／返る／孵る／かえる
- あう／会う／合う／逢う／遭う
- 結う／ゆう／いう／云う／謂う
- 買える／替える／かえる／飼える／代える／換える
- 揚る／揚がる／騰がる／騰る／あがる／上がる
- 撃ち取る／討取る／討ちとる／打ち取る／討ち取る／打取る
- 篤い／熱し／厚い／暑し／あつい／熱い／あつし／暑い／厚し

また逆に、下記のような語はまとめ上げがされていない。

- 空缶／空カン／空かん／あき缶／空きカン／空き缶
- 抱合わせ／抱きあわせ／抱き合せ／抱き合わせ
- 立ちあがり／立ち上がり／立上がり／立ち上り
- 独り暮し／一人暮し／一人暮らし／独り暮らし
- あかとんぼ／赤とんぼ／赤トンボ
- シューマイ／シウマイ／焼売／シュウマイ

3.2 UniDic

UniDic は国立国語研究所のコーパス構築に使用されている形態素解析用辞書である。UniDic には「語彙素」という概念があり、この情報を用いて表記ゆれのまとめあげを行っている。しかし、特にひらがな語と漢字語間のまとめあげについて問題があるように見える。例えば、「にる」という単語は対応する漢字表記との間で表記ゆれが存在する。しかし、「にる」の場合は「煮る」「似る」など、複数の対応する漢字があるため、この場合に表記ゆれを吸収することは語義曖昧性解消することを意味する。しかし、我々の見る限り、UniDic を用いて形態素解析すると「にる」（例えば「母親ににる」）は常に「煮る」（語彙素表記で、母親／に／煮る）となり、語義曖昧性解消が出来ているようには見えない。この結果、UniDic では誤った表記ゆれの解消を行ってしまう。このような例は「にる」だけでなく、対応する漢字表記が複数存在するほとんどのひらがな表記に対して発生するようである。

4 雪だるまプロジェクト

我々は表記ゆれを含む形態素解析の諸問題に対処するため、2015 年から雪だるま（英語名 SNOWMAN）というプロジェクトを立ち上げ、新たに同名の日本語単語解析システムを作成している。この単語解析器は、いわゆる形態素解析器の処理後に、我々の構築した表記統制辞書を用いて表記ゆれの集約を行う。その後に、形態素結合を行うことで形態素から単語への変換処理を行い、最後にこれら単語列に対して同義語辞書 [2] を用いて各単語に同義性情報の付与を行う。我々は、形態素解析を含む以上の処理全体を「単語解析」と呼び、従来の形態素解析との違いを明確にしている。以上の処理の全体像を図 1 に示す。表記ゆれ以外の問題については文献 [3] を参照してほしい。

現在我々は UniDic を対象に表記ゆれの集約作業を行っている。この作業は UniDic や NAIST-jdic など他の表記ゆれ辞書の情報を統合させるだけでなく、問題のある表記ゆれ集約を削除するなどして日本語の表記ゆれ辞書として最大規模かつ最高品質とすることを目指して手作業を続けている。

これまでの作業で、活用情報を除いた UniDic 異なり語数 319,951 語を 287,058 語に集約することができた。これは概ね 10% の語彙数減少である。また、頻度別にみると、高頻度語（Web 日本語 N グラムにおける頻度 1 万以上の語）は、UniDic 中に 116,353 語あるが集約の結果 98,611 語になり、低頻度語（同言語資源で出現頻度 20 未満の語）は 53,850 語が表記ゆれ集約の結果 51,072 語となった。以上から、集約できた語彙は高頻度のものから低頻度のものまで様々であり、頻度に関わらず幅広く集約されていることが推察される。

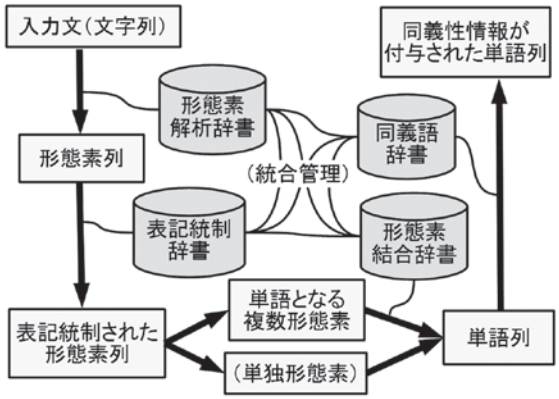

図 1　日本語解析システム「雪だるま」の単語解析部の処理の流れ

表記ゆれの吸収、すなわち語彙の集約によって同一の語数であっても網羅性は向上する。Web 日本語 N グラムの頻度情報を用いて我々の辞書と UniDic を調査した結果を図 2 に示す。これによると、語彙数を少なくすればするほど両者の差異は大きく、例えば高頻度 10,000 語で網羅性を見ると、UniDic では 96.8% であるが、語彙の集約を行った我々の辞書では 97.5% となり、0.7 ポイントの向上となった。また、図 2 から分かるように我々の辞書で高頻度 2,500 語（二つ目の測定点）を用いれば 91.6%、5,000 語（三つ目の測定点）で 95.1% の出現語彙を被覆することになり、少ない語彙数でかなりの出現を網羅できていることが分かる。

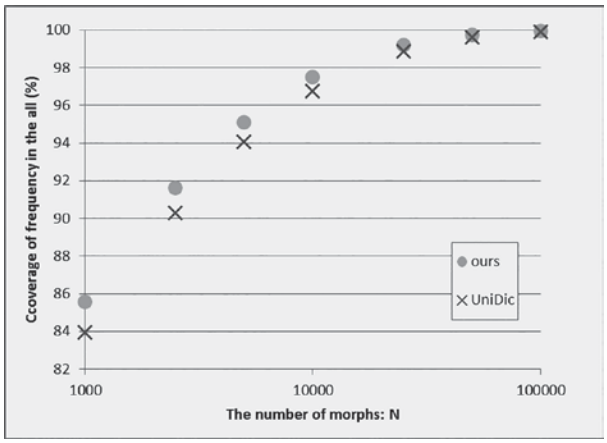

図 2　単語数とテキストの語彙網羅率の関係

5 おわりに

本稿では自然言語処理における表記ゆれ問題について議論し、我々の取り組みについて紹介した。表記ゆれは地味な現象であるが本稿で議論してきたようにその影響は小さくない。これに対し、一つ一つ表記ゆれ情報を言語資源化していくことで着実に精度は向上していく。このように、自然言語処理を本気で精度向上させるには、「地味な現象」に対して業績にならない「地味な努力」が必要である。

なお、雪だるまプロジェクトの URL は下記の通りで

ある。今後システムの公開や情報提供を当該サイトで行っていく。

雪だるまプロジェクトURL
http://snowman.jnlp.org/

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、課題名「日本語教育用テキスト解析ツールの開発と学習者向け誤用チェッカーへの展開」の助成を受けています。

参考文献

[1] 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, 2012

[2] Kazuhide Yamamoto and Kanji Takahashi. Construction of Japanese Semantically Compatible Words Resource. Proceedings of the International Conference on Asian Language Processing (IALP 2015) , 2015.

[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 (IALP 2015) , 2015.