4B1-4

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

Normalization of Text in Microblogging Based on Machine Learning

佐々木 彬[*1] Akira SASAKI　水野 淳太[*2] Junta MIZUNO　岡崎 直観[*1*3] Naoaki OKAZAKI　乾 健太郎[*1] Kentaro INUI

[*1]東北大学 Tohoku University　[*2]独立行政法人情報通信研究機構 National Institute of Information and Communications Technology(NICT)

[*3]科学技術振興機構　さきがけ Japan Science and Technology Agency(JST)

In recent years, microblogging such as Twitter has become increasingly popular. As a result, related research has been increasing in the field of natural language processing. However, there are many disadvantages because of the various kinds of spoken language and internet slang in microblogging. Therefore, this study was made to normalize text in micloblogging.

1. はじめに

Twitter をはじめとするマイクロブログの利用者数は近年爆発的に増加し，個人だけでなく企業による情報発信や意見交換のためにも活用されている．これに伴いマイクロブログを対象とした研究も増加しており，中でも自然言語処理分野では，多くの研究成果が報告されている．その中には，企業による報告も少なくない．企業は，マイクロブログ上の多くのユーザの意見，口コミなどを分析することにより，自社製品の改善などに活かすことができる．

しかしながら，マイクロブログ上のテキストには話し言葉やスラングが多く含まれるため，既存の解析器では解析誤りが生じる場合がある．例として，(1) のテキストを考える．

(1) A 社のケータィに変えたーーー使いにくぃー＞＜；

(1) のテキストは，A 社の携帯電話を利用しているユーザの不満であると解釈できる．この文に対して，言語解析を行うと，「ケータィ」というくだけた表記や,「変えた」に続く長音記号，文末の顔文字などにより，解析誤りが生じてしまう．このような解析誤りを防ぐために，元のテキストを解析器が想定しているテキストに変換することを考える．例えば，形態素解析器の MeCab [Kudo 04] に対しては，学習データとなっている新聞記事の表記に準じた形式に変換することによって，解析誤りを減らせることが期待できる．変換の基本方針は，日本語の正書法に則り，現代仮名遣いへの変換（例：わたしぁ→わたしは）などを行う．正書法では考慮されていないが，アスキーアートなどの除去も行う．ただし，送り仮名の修正や，規範文法からの逸脱などは，本研究では取り扱わない．本研究ではこのような変換を”正規化”と呼ぶ．機械学習に基づく手法を提案する．先行研究では，ルールに基づく正規化が取り組まれてきたが，本研究では機械学習に基づく正規化手法を提案し，マイクロブログのテキストに対して適用した結果について報告する．

連絡先: 佐々木彬，東北大学情報科学研究科システム情報科学専攻，宮城県仙台市青葉区荒巻字青葉 6-3-09，022-795-7140，022-795-4285，aki-s@ecei.tohoku.ac.jp

2. 関連研究

英語を対象とした正規化の既存研究として，[Brody 11] は，coooool，coollll，cool というように文字の連接を 1 文字にすると同じ文字列になるような単語から辞書を作成し，それにより正規化を行うという手法を提案した．また，[Bo 11] は文字，音素に着目して，正規化前の単語が未知語であるとき，それと編集距離の短い既知語を対応させる辞書を構築した．

ここで，英語は単語の区切りがスペースで明示されているため，少なくとも各単語の判別は容易である．しかしながら，日本語の場合は英語の場合と異なり，単語の区切りが明確ではないため，形態素解析を行う必要がある．この際に，顔文字やインターネットスラングなどによって，単語分割が正しく行えない可能性が高い．そのため，単語が正確に区切られていることを前提としている，英語を対象とした手法を日本語に直接用いることはできない．

日本語を対象とした正規化の既存研究として，[Ikeda 09] は，はじめに人手による少数の正規化ルールを導入し，それらを結合することにより複雑なルールを構築するという手法を提案している．また，[Kudo 12] は「京都」を「きょうと」と表記するようなひらがな交じりの文に対して，既存の形態素解析器では解析誤りが生じることに着目し，それに有効な形態素解析手法を提案している．形態素解析器 JUMAN [Kurohashi 94] は，長音記号や小書き文字の含まれるテキストの形態素解析を行う際にはそれらを除いた表記を用いて辞書の検索を行う，などといった対処をしている．

しかしながら，マイクロブログ上のテキストには記号や顔文字などが含まれるものも多い．例えば，(2) のテキストには，2 文字目には不必要な読点が含まれ，末尾には顔文字が含まれる．

(2) ゆ、れたね (;´ Д ‘A

このように，マイクロブログ上のテキストにはインターネットスラング，記号や顔文字などの多様な表現が含まれている．しかしながら，それらの表現の種類は膨大であるため，個別のルールを人手により構築することは難しい．

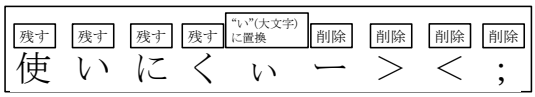

図 1: 1 文字単位のラベル付与の例

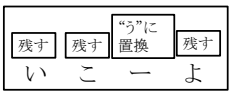

図 2: 訓練データ作成の例

3. 提案手法

人手によるルールの構築が難しいことから，本研究では機械学習に基づく正規化手法を提案する．形態素解析前の単語の区切りが明確ではない文を対象として，仮名遣いの修正などに対応するために，1 文字単位での正規化を行う．

3.1 文字単位の正規化ラベル付与

まず，正規化前のテキストに対して，それを正規化する際に，文字単位で付与するラベルについて定義する．ラベルは，残す，削除，置換の 3 種類である．それぞれの定義を以下に示す．

残す その文字に対して操作を加えず，そのまま残す際にこのラベルを付ける．

削除 その文字を削除する際にこのラベルを付ける．

置換 その文字を他のある文字に置換する際にこのラベルを付ける．また，置換後の文字列についても付記する．

ラベルの付与例を図 1 に示す．例では，小書き文字の「ぃ」は「い」に置換し，長音記号および顔文字は削除されるようラベルが付与されている．これらのラベルにしたがって正規化した結果は,「使いにくい」となる．これらのラベルを自動的に付与することができれば，文を正規化することができる．本研究では，ツイートデータに対して人手でラベルを付与し，それを教師データとして機械学習を行い，文字単位で正規化する．

3.2 訓練データ作成

まず，正規化前のテキストであるツイートデータを文単位に区切り，MeCab により形態素解析を行う．形態素解析結果は，正規化された文字を含む形態素の品詞情報を利用するために用いる．

例として，(3) の文を考える．

(3) いこーよ

この文に対して MeCab により形態素解析を行うと以下のようになる．

```
いこ　　動詞,自立,*,*,五段・カ行促音便,未然ウ接続,いく,イコ,イコ,いこ/逝こ,
ー　　名詞,一般,*,*,*,*,*
よ　　助詞,終助詞,*,*,*,*,よ,ヨ,ヨ,,
```

ここで，2 文字目の長音記号は，実際には「う」が適切であるため,「う」に置換するためのラベルを付与する．また，その他の文字についてはそのまま残すのが適切であるため，残すためのラベルを付与する．ラベル付与をすると，図 2 のようになる．

以上のようにして，形態素解析により品詞付けをした文と，それを人手によって正規化した文を得ることができる．以下では，人手によって正規化された文を「正解テキスト」と呼ぶ．これらの文の対を訓練データとすることで，どの文字がどの文字に置換されやすいか，どの文字が削除されやすいか，といった情報を学習させることができる．

3.3 訓練データの仕様

人手による訓練データ作成の際の仕様として，本研究では以下のように定める．基本方針としては，日本語の正書法に則ったテキストへの正規化を行う．しかしながら，正書法では考慮されないもののマイクロブログ上に多く見られ，なおかつ削除する必要のある表記が存在するため，それらも正規化の対象とする．

3.3.1 句読点

句読点については，過剰に繰り返されている場合，過剰な分のみ削除のラベルを付ける．例えば，(4) の文の末尾の句点は，初めの一つを除き削除のラベルを付ける．

(4) 心配だ。。。。

また，文中に不要な句読点が含まれる場合，削除のラベルを付ける．例えば，(2) の文中の「ゆ、れたね」の部分に含まれる読点には，削除のラベルを付ける．

3.3.2 挿入文字

不要な文字が挿入されている場合，それらの文字に削除のラベルを付ける．また，感嘆符，疑問符，その他の記号についても同様に，不要であれば削除のラベルを付ける．本研究では，感嘆符や疑問符は，正規化後の解析で利用されないことを想定している．利用する場合には，句読点の場合と同様に取り扱う．

例として，(5) の文を考える．

(5) なりたいですううううう！

この文において，”ううううう”の部分は特に意味が無い文字列であると考えられるため，各文字に削除のラベルを付ける．また，末尾の感嘆符についても削除のラベルを付ける．

3.3.3 アスキーアート

顔文字のようなアスキーアートが含まれる場合，アスキーアート内の文字に削除のラベルを付ける．例えば，(2) の文の末尾の顔文字に含まれる文字には，全て削除のラベルを付ける．

3.3.4 文字の置換

既存単語からの文字の削除，挿入，置換によって生じた派生単語については，元の単語へと戻すように正規化を行う．例として，(6) の文を考える．

(6) おはよおー

(6) において，4 文字目の”お”は,「おはよう」という単語の 4 文字目の”う”から派生したものと判断できる．また，5 文字目の長音記号は,「おはよう」という単語の末尾に挿入されていると判断できる．したがって，4 文字目には”お”から”う”への置換のラベルを付け，5 文字目の長音記号には削除のラベルを付ける．

置換については，”お”から”ぉ”のような小書き文字への置換，”し”から”U”のような見た目の類似した別の文字への置換がされている場合も考慮して正規化を行う．また，本研究では 1 文字から 1 文字への置換のみを対象とし，1 文字から複数文字への置換は対象外とする．これは，3.1 節で述べた置換のラベルについて，1 文字から 1 文字への置換にのみ対応するためである．

3.3.5 未知語

顔文字やアスキーアート以外の未知語については，既存単語に対する文字単位の編集により生じた派生単語以外は，正規化の対象外とする．例として，(7) の文の「さげぽよ」という未知語については，既存単語の派生語ではないことから，文字単位の編集では対処できないため，正規化の対象外とする．

(7) 朝からさげぽよです

3.4 機械学習手法

文字単位での正規化を，系列ラベリングの問題と考えて，CRF(Conditional Random Fields) [Lafferty 01] を用いて学習させる．また，CRF の実装として CRFsuite [Okazaki 07] を用いる．

3.5 素性

本手法で用いる各素性について述べる．ここで，例として (8) のテキストを考える．

(8) いこーよ！

周辺文字

文字の前後数文字までを素性とする．

母音

文字が母音であるか否かを素性とする．ここで，ひらがなの”あいうえお”，カタカナの”アイウエオ”を母音とする．例えば，例に挙げたテキストの文字”い”については母音素性は True となり，文字”こ”については母音素性は False となる．

品詞

形態素解析の結果，文字に付与された品詞を素性とする．例に挙げたテキストを形態素解析すると以下のようになる．

```
いこ    動詞,自立,*,*,五段・カ行促音便,未然ウ接続,いく,
イコ,イコ,いこ/逝こ,
ー      名詞,一般,*,*,*,*,*
よ      助詞,終助詞,*,*,*,*,よ,ヨ,ヨ,,
！      記号,一般,*,*,*,*,！,！,！,,
```

形態素解析の結果，文字列”いこ”には動詞という品詞が付与されている．このとき，文字”い”の品詞素性は動詞となる．

文字種

文字の文字種を素性とする．ここで文字種は，ひらがな，カタカナ，漢字，アルファベット，その他，とする．

4. 実験

提案手法により実際にマイクロブログ上のテキストの正規化を行えるかを評価した．

4.1 評価尺度

評価尺度としてレーベンシュタイン距離 (編集距離) を用いる．ここで，削除，挿入，置換のコストを 1 とする．この評価尺度を用いて，訓練データに含まれるテキストについて，モデルによる正規化の前後で，正解テキストとの距離が近づくか，遠ざかるかを見る．本手法を評価するにあたって，機械学習を用いない 2 種類のベースラインを設定した．

表 1: 正解テキストからの平均距離

比較対象	平均距離
正規化前	0.8770
ベースライン 1 による正規化後	0.7866
ベースライン 2 による正規化後	0.7657

ベースライン 1

ベースライン 1 では，同じ文字が連続する場合，その文字列全体を削除する，というルールを適用する．例として，(9) の文を考える．この文をこのルールで正規化すると (10) のようになる．

(9) やばあああああああああいwwwww

(10) やばい

ベースライン 2

ベースライン 2 では，同じ文字が連続する場合，その文字列を 1 文字を除き削除する，というルールを適用する．(9) の文をこのルールで正規化すると (11) のようになる．

(11) やばあいw

4.2 実験設定

実験の際には株式会社ホットリンクより提供された，ツイートデータを用いる[*1]．このツイートデータには，2011 年 3 月 11 日から 2011 年 3 月 29 日までの約 2 億 1 千万のツイートが含まれる．ここから無作為に抽出した 1000 ツイートを人手によりラベル付けし，半数の 500 ツイートを訓練データに，もう半数の 500 ツイートをテストデータに用いる．1000 ツイートは 1495 文からなり，訓練データの 500 ツイートには 731 文，テストデータの 500 ツイートには 764 文が含まれていた．

ここで，英語や韓国語などの日本語以外の言語の文については，本手法の対象ではないため，あらかじめ削除している．また，日本語の文についても，文中に含まれる URL や,「RT」などの Twitter 特有の記号は削除している．系列ラベリングに用いるラベルとしては，3.1 節で述べた，残す，削除，置換の 3 種類のラベルを用いる．

4.3 正規化種類の分布

元のツイートの各文字について人手により正規化を行なった結果，削除の対象となった文字は 15012 文字，置換の対象となった文字は 31 文字含まれていた．削除の対象となった文字とその頻度を表 3 に示す．削除の対象となった文字のうち，顔文字に含まれる文字数は 251 文字であった．

4.4 実験結果

モデルによりテストデータ中の各テキストを正規化し，正解テキストとの距離を比較した．はじめに，正規化前のテキスト，ベースライン 1 による正規化後のテキスト，ベースライン 2 による正規化後のテキストと，正解テキストの平均距離を表 1 に示す．これより，ベースライン 1 とベースライン 2 による正規化により，正規化前のテキストを正解テキストに近づけられていることが確認できた．これを踏まえ，素性を変えた各モデルによる結果を表 2 に示す．

まず，正規化の対象となる文字の周辺何文字まで考慮すべきかを調べるため，1 文字から 5 文字まで変化させたときの平均距離を表 2 の最初の 5 行に示す．その結果，前後 3 文字を考

*1 http://www.hottolink.co.jp/press/936

表 2: 素性を変化させたときの平均距離

素性	平均距離
前後 1 文字	0.3796
前後 2 文字	0.4188
前後 3 文字	0.3691
前後 4 文字	0.4672
前後 5 文字	0.4463
前後 3 文字, 母音	0.3469
前後 3 文字, 母音, 品詞	0.4267
前後 3 文字, 母音, 文字種	0.5406
前後 3 文字, 母音, 品詞, 文字種	0.5576

表 3: 削除のラベルを付与された文字と頻度 (上位 10 文字)

文字	頻度
！	2,376
w	2,070
・	1,107
…	1.098
？	873
w	702
あ	684
ー	603
！	369
^	270

慮する場合の平均距離が最も短くなった．以降では，この素性を中心に他の素性を加えていき，平均距離の変化を調べる．

前後 3 文字の素性に，母音の素性を加えた結果を表 2 の 6 行目に示す．前後 3 文字の素性のみを使う場合と比較して，平均距離を短くすることができた．母音の素性が有効であると考えた理由は，マイクロブログのテキストには (9) のように母音が連接されるようなテキストが多く含まれるためである．前後 3 文字の素性に母音の素性を加えた結果，性能はより向上した．

次に，前後 3 文字の素性，母音の素性に加えて，品詞の素性，文字種の素性を組み合わせて実験を行なった．しかしながら，その際には性能は悪化した．そのため，元のテキストに対して形態素解析を行なった際の品詞や，文字種は，文字の削除や置換のされやすさには直接は関係しないのであると考えられる．

また，正規化が必要なテキストの中でも，モデルによって正規化が行えなかったテキストも存在した．これの理由として，訓練データ不足ということが第一に挙げられる．先述の通り，マイクロブログ上のテキストには多様な表現が含まれる．本実験では 500 ツイートのみを訓練データとして用いたが，これらのツイートに含まれるテキストにはあくまでその多様な表現の一部しか含まれない．訓練データを増やせば，それに比例して対応できるマイクロブログ上のテキストの表現は増加し，より多くのテキストを正規化できるようになると考えられる．本手法では無作為に抽出したテキスト集合を全て人手で確認しながらラベル付けすることで訓練データを作成していたが，正規化の必要な表現を頻繁に使うユーザのツイートを優先的に見る，などの方法で正規化が必要となるようなテキストのみをあらかじめ抽出することができれば，訓練データの作成が容易になると考えられる．

5. おわりに

本研究では，マイクロブログ上のテキストを対象として，正書法から逸脱した仮名遣い，派生語，アスキーアートなどを正規化する手法を提案した．今後の課題として，ラベルや素性の見直し，訓練データの拡充が挙げられる．また，今回マイクロブログ上のテキストとして Twitter におけるツイートデータを対象としたが，これには発信したユーザの情報も付与されている．マイクロブログ上では口語表現，インターネットスラング，顔文字の含まれるテキストを頻繁に発信するユーザとそうでないユーザに分かれると考えられるため，訓練データの作成や素性の設定の際にユーザ情報を考慮することは有用であると考えられる．

参考文献

[Bo 11] Bo Han, and Timothy Baldwin. ”Lexical normalisation of short text messages: Makn sens a# twitter.” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Vol. 1. 2011.

[Brody 11] Samuel Brody, and Nicholas Diakopoulos. ”Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs.” Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[Ikeda 09] 池田和史, 柳原正, 松本一則, 滝嶋康弘. ”ブログ的表記を正規化するためのルール自動生成方式の提案と評価.” 日本データベース学会論文誌 8.1 (2009): 23-28.

[Kudo 04] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. ”Applying conditional random fields to Japanese morphological analysis.” Proceedings of EMNLP. 2004.

[Kudo 12] 工藤 拓, 市川 宙, David Talbot, 賀沢秀人. Web 上のひらがな交じり文に頑健な形態素解析, 言語処理学会全国大会 NLP-2012, 2012

[Kurohashi 94] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. ”Improvements of Japanese morphological analyzer JUMAN.” Proceedings of The International Workshop on Sharable Natural Language. 1994.

[Lafferty 01] John Lafferty, Andrew McCallum, and Fernando CN Pereira. ”Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” (2001).

[Okazaki 07] Naoaki Okazaki. ”CRFsuite: a fast implementation of conditional random fields (CRFs).” URL http://www.chokkan.org/software/crfsuite (2007).

[Stenetorp 12] Pontus Stenetorp, Sampo Pyysalo, Goran Topi, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. ”BRAT: a Web-based Tool for NLP-Assisted Text Annotation.” EACL 2012 (2012): 102.