2012年度　卒 業 論 文

系列ラベリングによる
マイクロブログ上の文の正規化

2013年3月31日

情報知能システム総合学科
(学籍番号: A9TB2096)

佐々木 彬

東北大学システム工学部

概 要

近年，Twitter 等のマイクロブログを対象とした自然言語処理関連の研究が増加している．しかしながら，マイクロブログには，ブログ特有の表現やインターネットスラング，口語表現が入り交じっているため，基本的な自然言語処理である形態素解析さえ失敗するような文も多く含まれ，その後の自然言語処理へと悪影響が及ぶ場合がある．本研究では，マイクロブログを対象として何らかの自然言語処理を行う前処理として，系列ラベリングを用いて文の正規化を図る．

目 次

第1章　はじめに

1.1　背景

Twitter 等のマイクロブログの利用者数は近年爆発的に増加し，個人や企業により，情報発信や交流のために用いられている．また，2011 年の東日本大震災の際には，安否確認，避難情報などの重要な情報がマイクロブログ上に集まり，震災に関する情報源のひとつとしてマイクロブログは大きな役割を担った．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関連の研究は特に盛んに行われている．しかしながら，マイクロブログを対象とするにあたって，自然言語処理に通常用いられていた手法を適用できない場合があり，これにより不都合が生じる場合がある．以下に，マイクロブログ上のテキストの例を示す．

- まだまだおわらにゃい (＞＜*) ノ
- あゆたんキタ━━━━ (゜∀゜) ━━━━!!ww

以上のようなテキストには顔文字やインターネットスラング，口語表現が含まれ，自然言語処理における基本的な処理の形態素解析さえも失敗する場合がある．形態素解析に失敗してしまうと，例えば名詞の誤った認識などにより，その後の自然言語処理においても性能が落ちるなどの悪影響を及ぼす．本研究では，マイクロブログ上のこれらの自然言語処理に適さないテキストに着目する．

1.2　目的

本研究では，マイクロブログ上のテキストを自然言語処理に適した形へと正規化することを目的とする．例えば，以下のようなテキストを考える．

- 逃げたいお＞＜。 1 人怖いお (*｀ω´) だれか来てー・゜･ (＞_＜) ・゜･ 。

このようなテキストについて，Mecab[1] を用いて形態素解析を行うと以下のようになる．

```
逃げ 動詞,自立,*,*,一段,連用形,逃げる,ニゲ,ニゲ,にげ/逃げ,
たい 助動詞,*,*,*,特殊・タイ,基本形,たい,タイ,タイ,,
お 感動詞,*,*,*,*,*,お,オ,オ,,
＞ 記号,括弧閉,*,*,*,*,＞,＞,＞,,
＜ 記号,括弧開,*,*,*,*,＜,＜,＜,,
。 記号,句点,*,*,*,*,。,。,。,,
1 名詞,数,*,*,*,*,1,イチ,イチ,,
人 名詞,接尾,助数詞,*,*,*,人,ニン,ニン,,
怖い 形容詞,自立,*,*,形容詞・アウオ段,基本形,怖い,コワイ,コワイ,こわい/コワい/怖い/恐い,
お 接頭詞,名詞接続,*,*,*,*,お,オ,オ,,
( 記号,括弧開,*,*,*,*,"(","(","(",,
* 記号,一般,*,*,*,*,*,*,*,,
｀ 記号,一般,*,*,*,*,｀,｀,｀,,
ω 記号,アルファベット,*,*,*,*,ω,オメガ,オメガ,,
´ 記号,一般,*,*,*,*,´,´,´,,
) 記号,括弧閉,*,*,*,*,")",")",")",,
だれ 名詞,代名詞,一般,*,*,*,だれ,ダレ,ダレ,,
か 助詞,副助詞／並立助詞／終助詞,*,*,*,*,か,カ,カ,,
来 動詞,自立,*,*,カ変・来ル,連用形,来る,キ,キ,き/来,
て 助詞,接続助詞,*,*,*,*,て,テ,テ,,
ー 名詞,一般,*,*,*,*,*
・ 記号,一般,*,*,*,*,・,・,・,,
゜゜ 名詞,サ変接続,*,*,*,*,*
・ 記号,一般,*,*,*,*,・,・,・,,
( 記号,括弧開,*,*,*,*,"(","(","(",,
＞ 記号,括弧閉,*,*,*,*,＞,＞,＞,,
\_ 記号,一般,*,*,*,*,\_,\_,\_,,
＜ 記号,括弧開,*,*,*,*,＜,＜,＜,,
) 記号,括弧閉,*,*,*,*,")",")",")",,
・ 記号,一般,*,*,*,*,・,・,・,,
゜゜ 名詞,サ変接続,*,*,*,*,*
・ 記号,一般,*,*,*,*,・,・,・,,
。 記号,句点,*,*,*,*,。,。,。,,
EOS
```

このように，顔文字や不自然な語尾が含まれるようなテキストはそのままの形では形態素解析に失敗してしまう．そこで，以下のように元のテキストを人手によって正規化する．

- 逃げたい。1 人怖い。だれか来て。

このテキストに対して再度形態素解析を行うと以下のようになる．

```
逃げ 動詞,自立,*,*,一段,連用形,逃げる,ニゲ,ニゲ,にげ/逃げ,
たい 助動詞,*,*,*,特殊・タイ,基本形,たい,タイ,タイ,,
。 記号,句点,*,*,*,*,。,。,。,,
1 名詞,数,*,*,*,*,1,イチ,イチ,,
人 名詞,接尾,助数詞,*,*,*,人,ニン,ニン,,
怖い 形容詞,自立,*,*,形容詞・アウオ段,基本形,怖い,コワイ,コワイ,こわい/コワい/怖い/恐い,
。 記号,句点,*,*,*,*,。,。,。,,
だれ 名詞,代名詞,一般,*,*,*,だれ,ダレ,ダレ,,
か 助詞,副助詞／並立助詞／終助詞,*,*,*,*,か,カ,カ,,
来 動詞,自立,*,*,カ変・来ル,連用形,来る,キ,キ,き/来,
て 助詞,接続助詞,*,*,*,*,て,テ,テ,,
。 記号,句点,*,*,*,*,。,。,。,,
EOS
```

このようにして，元のテキストを正規化することによって，形態素解析の失敗を防ぐことができる．本研究では，マイクロブログ上の自然言語処理に適さないテキストについて正規化するシステムを構築することを目的とする．

第2章　関連研究

マイクロブログやブログ上の表記の正規化に関して，英語を対象とした既存研究，日本語を対象とした既存研究がある．

2.1　英語を対象とした正規化

英語のマイクロブログ上の正規化が必要となるようなテキストには下記のようなものがある．

- He is coooooooooooooolll

ここで，英語の場合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り，例えば上記の例の coooooooooooooolll は一つの単語であると判断することができる．Brody ら [2] は，cooooool，cooollll，cool というように文字の連接を削ると同じ文字列になるような文字列から辞書を作成し，それにより正規化を行うという手法を用いている．

2.2　日本語を対象とした正規化

日本語のマイクロブログ上の正規化が必要となるようなテキストには下記のようなものがある．

- 彼はかっこいいいいいいいい

日本語テキストの場合は英語テキストの場合と異なり，単語の区切りが明示されていないため，形態素解析を行う必要がある，しかし，上記のテキストのように末尾に不要な文字が挿入されている場合，形態素解析に失敗してしまう．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については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手法としては，池田ら [3] の，少数の人手による正規化ルールを組み合わせて複雑なルールを生成するという手法がある．しかしながら，マイクロブログ上には人手によるルールを与えるのが困難なテキストが多数存在する．例えば，以下のようなテキストがある．

- 20 日からのバリ島楽しみだね (*＞ω＜*) あと五日！笑
- いやああぁぁたあすけてえぇえええええええええ
- (´ ° д ° ‘)(´ ° д ° ‘) あんのくそったれえええええええええええええ
- 無事でとりまよかた＞＜

以上のように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ため，人手によるルール作成は難しい．そのため，テキスト内の文字について 1 文字ずつ見て，それぞれ削除，置換などの編集をすることができれば望ましいと考えられる．そこで本研究では，系列ラベリングによるテキストの正規化手法を提案する．

第3章　提案手法

本研究では系列ラベリングによるテキストの正規化に取り組む．

3.1　文字単位のラベル付けによるテキストの正規化

正規化の際，はじめに入力として与えられたテキストに対して 1 文字単位でラベルを付ける．用いるラベルは以下の 3 種類とする．

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．ここで，どの文字に置換するかによって異なるラベルを付ける．例えば，ひらがなの”い”に置換する際には”い”に置換というラベルを付ける．

上記のラベルを用いた正規化の具体例を以下に示す．まず，正規化前のテキストとして以下のものを考える．

- おはよおぉおぉございます

次に，上記のテキストの「おはよおぉおぉ」の部分に対して 1 文字ずつ，図 3.1 のようにラベルを付けるとする．図 3.1 のようにラベルを付けることができたとすると，各文字に付けられたラベルを考慮して 1 文字単位で削除，置換の操作を行えば，正規化前のテキストは以下のように正規化することが可能となる．

- おはようございます

このラベルを全て人手で正確に付けることができれば高い精度で正規化を行えるが，膨大なデータを扱う際にも全て人手でラベル付けしてしまうと，非常にコストが高くなってしまう．本研究

図 3.1: 1 文字単位のラベル付けの例

では，人手でラベル付けしたテキストの集合を訓練データとして機械学習によりモデルを生成し，そのモデルにより系列ラベリングを行う手法を提案する．

3.2 訓練データ作成

テキストにラベル付けを行うために，アノテーションツールの brat[4] を用いる．brat の概観を図 3.2 に示す．ここで brat を用いる理由は，人手によるテキストへのラベル付けが容易であるためである．

訓練データ作成の際，はじめに元の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．ここでツイートデータを文単位に区切るのは，形態素解析の際に入力を 1 文単位とする必要があるためである．例として，以下の文を考える．

- いこーよ！

この文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,*,*,*,*,！,！,！,,
```

この形態素解析結果について，brat で読み込むと図 3.3 のようになる．ここで，図 3.3 の verb，noun，part，symb という各文字列は，形態素解析結果の動詞，名詞，助詞，記号にそれぞれ対応している．

次に，図 3.3 のラベルを人手により訂正すると，図 3.4 のようになる．訂正しているラベルは長音符”ー”に付けられている noun(名詞) のラベルで，人手によって aux(助動詞) のラベルへと訂正されている．また，実際にはこの文字は長音符”ー”ではなく”う”が適切であるため，注釈として適切な文字”う”を記入している．

以上のようにして，形態素解析によりラベル付けをした図 3.3 の文と，そのラベルを人手によって訂正した図 3.4 の文を得ることができる．このような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，などを，機械学習によりモデルを生成することが可能となる．

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．

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，以下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．

- 心配だ。。。。

その他の文字

その他の文字については，形態素解析に悪影響を与えてしまうような文字や，不要な文字には削除のラベルを付ける．例として，以下の文を考える

- 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，感嘆符”！”については，必要のない文字であると考えられるため，削除のラベルを付ける．加えて，顔文字の含まれる以下の文を考える．

- ゆれたね…(;´ Д ‘A

このような文については，顔文字に含まれる各文字についても削除のラベルを付ける．

3.4 機械学習手法

系列ラベリングの機械学習の手法として CRF(Conditional Random Fields)[5] を用いる．また，CRF の実装として CRFsuite[6] 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，例として以下のテキストを考える．

- いこーよ！

周辺文字

対象となる文字の前後数文字までを素性とする．例えば，例に挙げたテキストの 3 文字目の長音符”ー”について，前後 2 文字までを素性とすると，

```
2 文字前: い
1 文字前: こ
1 文字後: よ
2 文字後: ！
```

となる．

母音

対象となる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，対象となる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．

図 3.2: brat の概観

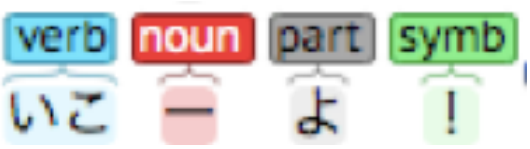

図 3.3: 訂正前のラベルの例

図 3.4: 人手による訂正後のラベルの例

```
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．

第4章　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) を用いた．ここで，削除，挿入，置換の各操作のコストを 1 とした．この評価尺度により，モデルによる正規化が成功しているかを評価した．以下に，この評価尺度による例を示す．

4.1.1　評価例

正規化前のテキストと人手による正規化後のテキスト，またモデルによる正規化後のテキスト 2 種類として以下のような例を考える．

正規化前のテキスト
　おはよううううございまつ

人手による正規化後のテキスト
　おはようございます

モデルによる正規化後のテキスト 1
　おはよううございます

モデルによる正規化後のテキスト 2
　うはよううううございまつ

まず，正規化前のテキストと人手による正規化後のテキストとのレーベンシュタイン距離を考えると，編集の過程は図 4.1 のようになり，レーベンシュタイン距離は 3 となる．

次に，モデルによる正規化後のテキスト 1 と人手による正規化後のテキストのレーベンシュタイン距離を考えると，編集の過程は図 4.2 のようになり，レーベンシュタイン距離は 1 となる．

最後に，モデルによる正規化後のテキスト 2 と人手による正規化後のテキストのレーベンシュタイン距離を考えると，編集の過程は図 4.3 のようになり，レーベンシュタイン距離は 4 となる．

これより，モデルによる正規化後のテキスト 1 と人手による正規化後のテキストのレーベンシュタイン距離は，正規化前のテキストと人手による正規化後のテキストのレーベンシュタイン距離と比較して短くなっていることがわかる．

また，モデルによる正規化後のテキスト 2 と人手による正規化後のテキストのレーベンシュタイン距離は，正規化前のテキストと人手による正規化後のテキストのレーベンシュタイン距離と比較して長くなっていることがわかる．

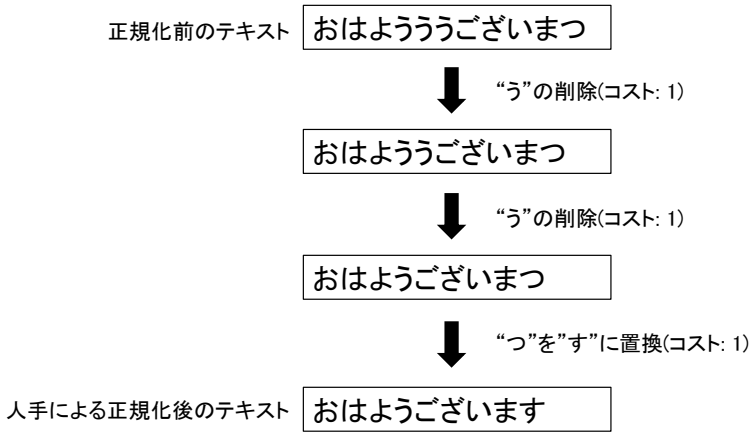

図 4.1: レーベンシュタイン距離の例 1

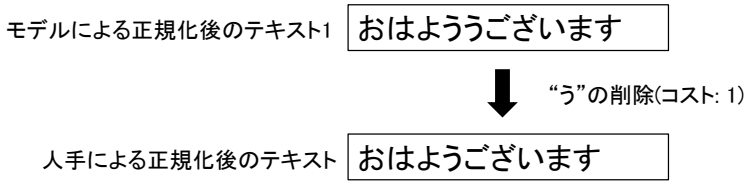

図 4.2: レーベンシュタイン距離の例 2

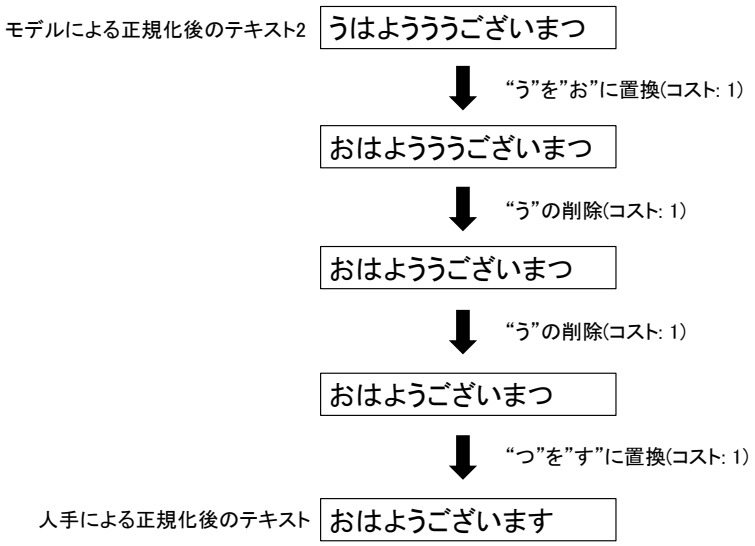

図 4.3: レーベンシュタイン距離の例 3

よって，モデルによる正規化後のテキスト 1 は正規化前に比べて人手により正規化したテキストに近づいていると判断できるが，モデルによる正規化後のテキスト 2 は正規化前に比べて人手により正規化したテキストから遠ざかっていると判断できる．

以上のように，実験ではレーベンシュタイン距離を用いることでモデルの性能を評価する．

4.2 ベースライン

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．

ベースライン 1

連続した同じ文字を削除するという手法をベースライン 1 として用いる．例として，以下の文を考える．

- やばあああああああああいwwwww

上記の文をこの手法で変形すると以下のようになる．

- やばい

ベースライン 2

連続した同じ文字を，1 文字を除き削除するという手法をベースライン 2 として用いる．例として，以下の文を考える．

- やばあああああああああいwwwww

上記の文をこの手法で変形すると以下のようになる．

- やばあいw

4.3 実験設定

実験の際には Hottolink 社より提供された，Twitter におけるツイートデータを用いる．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．これらのツイートデータ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，URL のみからなる文や，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．系列ラベリングに用いるラベルとしては，第 3 章で述べた，残す，削除，置換の 3 種類のラベルを用いる．

表 4.1: 各モデルによる正規化テキストと正解テキストとの距離

素性	正規化後のテキストと正解テキストとの平均距離
前後 1 文字まで	0.3796
前後 2 文字まで	0.4188
前後 3 文字まで	0.3691
前後 4 文字まで	0.4672
前後 5 文字まで	0.4463
前後 3 文字まで, 母音	0.3469
前後 3 文字まで, 品詞	0.4450
前後 3 文字まで, 母音, 品詞	0.4267

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，人手による正規化後のテキストとの距離を比較した．ここでまず，正規化前のテキストと人手による正規化後のテキストとの平均距離は 0.8770 であった．また，ベースライン 1 による変形後のテキストと人手による正規化後のテキストとの平均距離は 0.7866 となり，ベースライン 2 による変形後のテキストと人手による正規化後のテキストとの平均距離は 0.7657 であった．これを踏まえ，素性を変えた各モデルによる結果を表 4.1 に示す．以下，人手による正規化後のテキストを「正解テキスト」と呼ぶ．ここで，前後 3 文字を素性とした際の性能が最も優れていたため，それに母音の素性，品詞の素性を組み合わせたモデルもまた評価した．

4.5 分析

表 4.1 の通り，各モデルによる正規化後のテキストと正解テキストとの平均距離は，正規化前のテキストに比べて短くなった．そのため，生成したモデルは正規化の必要なテキストを正規化できていると考えられる．また，ベースライン 1 やベースライン 2 の単純な手法と比較して，モデルによる正規化の方がより距離を短くすることができた．これより，機械学習を用いた正規化は単純な手法による正規化よりも有用であると考えられる．次に，前後 3 文字の素性に加えて母音の素性を加えた際，わずかながら性能が向上した．これは，マイクロブログのテキストには以下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と考えられる．

- よかったあああああああああ
- やばあああああああああいwwwww
- なりたいですううううう！

このように，母音の素性を加えた際には性能の向上が見られたものの，品詞の素性を加えた際には性能が悪化した．これより，元のテキストを形態素解析した際の品詞は，その品詞を付けられた文字の削除，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．ここで，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，何らか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．

第5章　おわりに

機械学習により訂正モデルを生成し，系列ラベリングを適用することで，マイクロブログ上のテキストを自然言語処理に適した形へと近づけることができた．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．現時点ではマイクロブログ上のテキストの一部の正規化に留まっているが，より多くのテキストを正規化できるようなシステムを構築することができれば，マイクロブログ上のテキストを扱うための前処理として有用であると考えられる．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．この点についても，今後研究を進めていく上で取り組む必要がある．

謝 辞

本研究を進めるにあたり，ご指導頂きました乾健太郎教授，岡崎直観准教授に感謝致します．また，本研究について多くのご指摘を下さいました乾・岡崎研究室の皆様に感謝します．

参 考 文 献

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

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

[3] 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.

[4] 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.

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

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