マイクロブログを用いた音声認識用言語モデルの構築及び分析

市川博通，黒澤義明，目良和也，竹澤寿幸
広島市立大学大学院情報科学研究科

1. はじめに

近年, Twitter[1]等のマイクロブログが盛んに利用されるようになっている．Twitterとは，個々のユーザが「Tweet(ツイート)」と呼称される短文を投稿し，幅広いつながりを持てるコミュニケーションサービスである. Twitterは掲示板のような使い方，チャットのような使い方，blogのような使い方など様々な形態で利用されている．現在の入力はキーボードで行われているが，上肢障害の方や，手がふさがっている利用環境の場合，入力が困難である．そこで本稿では，音声入力による投稿に着目する．

音声入力を行うためには，音声認識技術が必要不可欠である．音声認識とは，人の話す音声言語をコンピュータによって解析し，話している内容を文字データとして取り出す処理のことである．音声認識エンジンとして現在，オープンソースの汎用大語彙音声認識エンジンJulius[2]が公開されている．Juliusで音声認識を行うには，音声の特徴を表す音響モデル，言葉のつながりを表す言語モデルが必要である．しかし，音声入力で投稿を行うことが出来ても，認識結果に誤りが起こるとユーザが期待していないTweetが投稿されてしまう可能性がある．認識誤りを低減させるためには，高精度の音響モデル，高精度の言語モデル，そして効率のよい認識エンジン(デコーダ)が必要になってくる．本稿では認識誤りを低減させるため，Tweetをベースとした言語モデルを構築し，性能評価を行った．

2. 現状における音声認識

近年のコンピュータ技術の発達と研究の進展によって，新聞記事等のあらかじめ用意されたテキストを読み上げる課題であれば，かなり高い精度で認識できる．しかしながら話し言葉においては，書き言葉で使用される語彙や言い回しが異なるため，大幅に認識性能が下がってしまう．今回音声認識の入力対象とするTwitterには略語，俗語，隠語，「なう」「だん」等のスラング，くだけた話し言葉等, 多くの言語形態を含んでいる. 従って，通常の音声認識で用いられている言語モデルではTwitterの投稿における音声認識で, 高い精度を達成することは出来ないと考えられる．

音声認識において，ドメインに適したテキストコーパスで学習した言語モデルを用いた方が，高い認識率が得られる[3]．そこでTwitter特有の表現をモデル化するため，本研究ではTweetを用いて言語モデルの構築を行った．言語モデル構築にあたり，信頼出来る統計量を得るため，大量なテキストコーパスを確保する必要がある．今回は独自に303,651件のTweetを収集した．本来言語モデルを構築する際に用いるコーパスは，人手で形態素解析したコーパスを用いる事が望ましい．そこで今回Tweetを自動で形態素解析した後，人手で後処理を行った．しかし，大量のTweetを人手で処理するには多大な労力がかかる．今回は人手で後処理を行うTweet数を10,000件とした．後処理を行う事により，言語モデルとして正しい統計が得られると期待できる．また，　本稿では10,000件のTweetを対象に, 形態素解析辞書に登録されていない場合は辞書に登録する作業を行った. 10,000件のTweetを用いた言語モデルの構築に加え，今回単語を登録した新しい辞書を用い，独自に収集したTweetを，自動的に形態素解析した言語モデルの構築も行った．

また, Twitterでは不特定多数のユーザが使用するため，認識できる語彙数はなるべく多い方が良いと考えられる．語彙数をカバーするため新聞記事，話し言葉のコーパスをTweetのコーパスに組み合わせて構築し，認識精度の向上を試みた．

3. 形態素解析と言語モデルの構築

単語の並びには出現しやすい並びと出現しにくい並びが存在する．このような単語の並びの出現確率のモデルは,「言語モデル」と呼ばれている．本稿では統計的言語モデルの中でも広く使われているN-gram言語モデルの構築を行った. N-gram言語モデルは発話が行われる前の単語列の生成確率を推定し，文の先頭にはどのような単語が出現しやすいか，また，ある単語の後ろにはどのような単語が出現しやすいかということを統計的に求めたモデルである. 一般的に日本語のテキストは，文が単語毎に区切られておらず，Tweetでも単語毎に区切られていない．そこで形態素解析を行い単語に分割する必要がある．

3.1. 形態素解析

本稿では形態素解析器として MeCab 0.97[4], 形態素解析辞書は mecab-ipadic-2.7.0-20070801 を用いて形態素解析を行った．しかし，Tweet では，形態素解析辞書に登録されていない単語(未知語)が多く含まれる．そこで 10,000 件の Tweet を対象に，解析できない単語については形態素解析辞書に登録し，形態素解析が行えるようにした．また，形態素としての区切りを誤って区切ってしまった場合，誤った並びに確率を与えてしまうため，形態素解析の後処理が必要になってくる．今回は形態素解析を行った後，人手で正しく区切られているか確認し，誤って区切ってしまった場合は正しい区切りに修正した．また，正しい読みで付与されていないと，発話が行われても単語列が生成出来ないため，形態素解析後の発音の確認も行った．また Tweet の抽出に関しては以下のような前処理を行った．

(1) 正規化

Tweet では全角と半角が混在するため，形態素解析辞書 mecab-ipadic に合わせて正規化を行った．また，数字表現の正規化を行うため，アラビア数字は漢数字に正規化した．表 1 が主な正規化である．

表 1　主な正規化

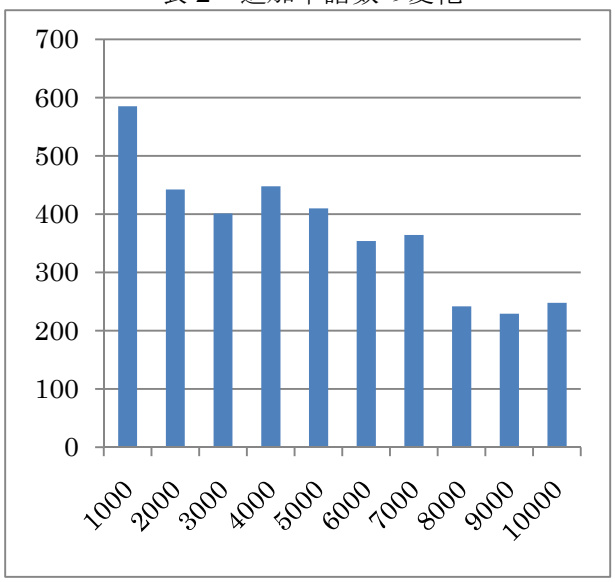
半角英字	→	全角英字
半角カタカナ	→	全角カタカナ
アラビア数字	→	漢数字
半角記号	→	全角記号

(2)文の分割

「。」,「.」,「！」,「？」,「♪」,「…」,「・」,「スペース」を文境界とし，Tweet を文単位に分割した．

10,000 件の Tweet を対象として, 形態素解析辞書に単語を登録した結果，1,000 件毎の Tweet における追加単語数の変化の様子を表 2 に示す．Tweet 数が増えていくにつれ，形態素解析辞書に登録する語数は減っていったが，10,000 件の Tweet だけでは，まだまだ Twitter で用いられている語を全て登録したとは言えない．10,000 件の Tweet により 3724 語の単語を新たに形態素解析辞書に登録した．登録した語は固有名詞，表記の揺れで起こる単語が多い傾向があった．

表 2　追加単語数の変化



3.2. 言語モデル構築と単語辞書

N-gram 言語モデルの作成ツールとして今回, Palmkit-1.0.31[5]を用いて前向き 2-gram と逆向き 3-gram の言語モデルを構築した．言語モデル平滑化手法としては Witten-Bell 法を用いた．語彙に対してはカットオフを行い，2 回以上出現した単語を基本語彙とした．

今回は独自に 303,651 件の Tweet を収集した．言語モデルとしてはまず, 10,000 件の Tweet を用いて言語モデルを構築した．また，辞書に単語の登録を行った新しい形態素解析辞書において，未知語を含んでいない 236,518 件の Tweet を用いて自動的に形態素解析を行い，言語モデルを構築した．

新しい形態素解析辞書の有効性を示すため，単語の登録を行う前の形態素解析辞書において，未知語を含んでいない 171,313 件の Tweet を用いて自動的に形態素解析を行い，言語モデルを構築した. また, 語彙数を増やすため 236,518 件の Tweet のコーパスに，新聞記事のコーパス[6]，話し言葉のコーパス[7]を組み合わせた言語モデルの構築も行った．

単語辞書は，認識対象とする単語とその読み(音素列表記)を定義している．Julius では，この単語を最少ユニットとして解探索を行う．単語辞書を構築する際に，複数の読みをもつ形態素は辞書に複数の読みを併記して登録を行った．また，同じ表記でも意味の違う形態素が存在するため，品詞情報も用いた．

表 3　実験に用いた言語モデル

言語モデル	語彙数
①:Julius ディクテーション実行キット付属の言語モデル[8]	65,250
②:日本語話し言葉コーパス(CSJ:Corpus of Spontaneous Japanese)[9]	27,249
③:10,000 件の Tweet から構築した言語モデル	5,337
④:単語の登録を行った形態素解析辞書を用いて構築した言語モデル	33,706
⑤:単語の登録を行う前の形態素解析辞書を用いて構築した言語モデル	29,758
⑥:④構築時に新聞記事 120 記事を組み合わせて構築した言語モデル	34,222
⑦:④構築時に新聞記事 1200 記事を組み合わせて構築した言語モデル	37,119
⑧:④構築時に新聞記事 12000 記事を組み合わせて構築した言語モデル	45,824
⑨:④構築時に話し言葉 200 対話を組み合わせて構築した言語モデル	33,966

表 4　①～⑤の言語モデルにおける実験結果

	①	②	③	④	⑤
単語正解率	66.31	55.80	63.78	76.13	71.57
単語正解精度	62.44	45.10	57.99	73.06	66.20
語彙数	60,250	27,249	5,337	33,706	29,758

4. 評価実験

4.1. 評価実験

評価実験では 10,000 件の Tweet を用いた言語モデルの有効性と，辞書に単語を登録した，新しい形態素解析辞書を用いて構築された言語モデルの有効性を確認するため，表 3 における言語モデルとの比較を行った．実験条件を以下に示す.

音声認識器：Julius
音響モデル：Julius ディクテーション実行キット付属の音響モデル.

テストセットとして，以下のデータを用いる.
音声データ：本研究室 14 人を対象
一人当たりの発話数：30 発話
計：420 発話

評価実験では言語モデルの評価を行うため，単語正解率，単語正解精度を求めて評価を行った．単語正解率，単語正解精度は次式で定義される.

$$\text{単語正解率} = \frac{\text{正解単語数}}{\text{全単語数}}$$

$$\text{単語正解精度} = \frac{\text{正解単語数}-\text{挿入誤り}}{\text{全単語数}}$$

また，漢字とかなの表記の曖昧性により，誤り率が増加してしまうため[10]，今回は機械的に判定した結果を確認し，表記の揺れの場合は正解とすることにした.

4.2. 実験結果

①~⑤の言語モデルについて, 実験により得られた評価の結果を表 4 に示す. 音声の内容としては, 書き言葉に近い発話も多くみられた．そのため既存の言語モデルである①，②でもある程度の単語正解率，単語正解精度が得られた.

また，今回人手で後処理を行った③の言語モデルでは，既存の言語モデル①と比べて単語正解率と単語正解精度は劣ったものの，②より性能が高い結果となった．少量の語彙数でも，ある程度の認識が行えたことから, Twitter で用いられる言い回しを効率よくモデル化出来たと考えられる.

新しい形態素解析辞書を用いて構築した④の言語モデルが単語正解率，単語正解精度ともに一番高い値となった. 実際の認識結果を見ても Twitter で使われている「なう」「だん」等，きちんと認識されていた．認識できなかった発話として，単語辞書に登録していない単語を含む発話が行われた場合，誤認識が見られた.

単語の追加を行う前の形態素解析辞書を用いて構築した⑤の言語モデルは単語の登録を行った④のモデルより性能が劣ったものの，既存の言語モデル①，②より単語正解率，単語正解精度ともに上回る結果となった.

表 5　⑥～⑨の言語モデルにおける実験結果

	⑥	⑦	⑧	⑨
単語正解率	74.85	74.52	73.78	75.94
単語正解精度	71.46	71.26	70.53	72.91
語彙数	34,222	37,119	45,824	33,966
形態素数	39,228	338,435	3,597,838	99,637

表 4 の結果から，人手で後処理を行ったコーパスを用いて構築した③の言語モデル，単語を登録した新しい形態素解析辞書を用いて構築した④の言語モデルの有効性が確認された．

次に⑥~⑨の言語モデルについて，実験により得られた結果を表 5 に示す．また④と組み合わせるコーパスの形態素数も示した．

⑥，⑦，⑧の言語モデルにおける結果から，新聞記事のコーパスを用いることで語彙数を増加させることが出来た．しかし，若干ではあるが新聞記事のコーパスの量を増やすほど，単語正解率，単語正解精度ともに低下していく傾向が見られた．

④の言語モデルと組み合わせて構築した言語モデルにおいては，⑨の言語モデルが単語正解率と単語正解精度が一番高い結果となった．この結果から，新聞記事と話し言葉のコーパスにおいては，話し言葉のコーパスの方が，Twitter の言い回しに適していると考えられる．

⑥，⑨の言語モデルの形態素数を見ると，⑥の方が形態素数が少ないのにもかかわらず，⑨の言語モデルよりも語彙数は増加していた．これは認識対象語彙を増加させた結果，混同しやすい単語の組が増加し，認識誤りが起きた可能性が考えられる．また，新聞記事では使われるが，Twitter ではほとんど使われない語彙を言語モデルとして与えてしまっていたとも考えられる．コーパスを組み合わせる場合において，新聞記事，Twitter の両方で使われる表現や語彙を選択することが望ましいと考えられる．

今回は小規模なテストセットでの実験であったが，単語を追加させた形態素解析辞書を行うことで，ある程度の認識精度が得られた．Twitter にはいろいろな言語が混在するが，それらを全て認識させようとすると，混合しやすい単語の組みが多くなると考えられる．

音声を書き起こしてみると，人によっては書き言葉に近い発話，話し言葉のような発話しか行わないという人もいた．人によって Twitter に投稿する言い回しや語彙も異なるため，ユーザ毎に適した言語モデルを構築することが望ましいと考えられる．

5. 今後の課題

テストセットとして作成した発話は計 420 発話という小規模な音声データであったため，正確な評価は行えていないという可能性もある．今後は更に音声の収集を行い，大量のテストセットを用いて評価実験を行っていきたい．また，今回は人手で形態素解析辞書の再構築を行ったが，今後は自動で再構築していく手法も検討していく．

謝辞

テストセットを作成するにあたり，協力して下さった研究室の皆さまに深く感謝いたします．

参考文献

[1]http://twitter.com/
[2]河原 達也，李 晃伸，“連続音声認識ソフトウェア Julius，”
[3]駒谷他，情処学論，vol.44，no.5，pp.1333-1342，2003.
[4]工藤 拓，山本 薫，松本 祐治，”Conditional Random Fields を用いた日本語形態素解析”情報処理学会研究報告，2004(47)，89-96，2004-05-13
[5]http://palmkit.souceforge.net/
[6]毎日新聞社．毎日新聞地方版．2005.
[7]翠他，情報処理学会研究報告.SLP，音声認識処理 2009(10)，39-44，2009-01-30
[8]http://julius.sourceforge.jp/index.php?q=juliuskit.html
[9] K.Maekawa ， H.Koiso ， S.Furui，and H.Isahara. Spontaneous Speech Corpus of Japanese. In Proceedings of LREC2000，pp.947-952，2000
[10]伊藤 克垣，山本 俊一郎，鹿野 清宏，中西 哲，ディクテーションにおける日本語の特徴を考慮した単語判定ツール．日本音響学会研究室発表会講演論文集，3-Q-19，春季 1999