D-027

マイクロブログサービスにおける自動リプライシステムの考案
Auto Replay System for Micro Blog Service

森下雄太 † 渋川翔太 ‡ 佐藤隆士 §
Yuta Morishita Shota Shibukawa Takashi Sato

1. はじめに

近年多くのユーザに利用されている Twitter[1] や Weibo[2] に代表されるようなマイクロブログサービスとは比較的短いテキストを発信し，ユーザ同士でコミュニケーションを行う SNS のひとつである．マイクロブログサービスでは投稿されたユーザの意見や考えに対し，別のユーザがコメントとしてそのユーザの意見を投稿できるリプライサービスが実装されている．そのコメントは発信者やコメントの投稿者，それらを閲覧するユーザにとって有用な情報をもたらすことも多い．

本研究ではユーザの投稿に対してリプライを自動的に付与することを目標とした研究となる．また，リプライに使用されるショートテキストは自動生成するわけではなく，過去に投稿された対話データ（投稿データとなる POST コーパスと返答となる CMNT コーパス）をコーパスとし，それらを対象に検索を実施し，適切と思われるショートテキストをリプライとして提示するものとなる．

評価型ワークショップ NTCIR-12 Short Text Conversation[3][4][5]（以下，STC）において提示されたリプライは表 1 に示す 2 つの評価点で評価される．

表 1: 2 つの評価点

評価	説明
coherent	POST とコメントに対する一貫性
useful	投稿者に対するコメントの有用性

コーパスは STC から提供された Twitter に投稿された対話データ 500,771 件（1,001,542Tweets）を対象に研究を実施した．

検索手法として元のショートテキストから検索語の抽出を行い，コーパスに対して検索を実施している．また，検索精度を向上させるために，品詞における検索語の重み設定やショートテキスト全体の特徴を示す特徴語の設定，ショートテキストのテキスト長に応じたスコアリングを行っている．

以下，本稿の構成について述べる．**2** では関連研究を述べる．**3** では STC から提供されたコーパスの分析結果について述べる．**4** では提案手法である検索語の抽出方法や検索語の重み設定，特徴語の設定方法，テキスト長に関するスコアリングについてを述べる．**5** では手案手法による実験結果について述べ，**6** で考察を行い，**7** で今後の展望について述べる．

†大阪教育大学大学院　総合基礎科学専攻
‡大阪教育大学　教育学部
§大阪教育大学　情報処理センター

2. 関連研究

マイクロブログ上での会話についても多くの研究や事例がある．例えば東日本大震災の際には Twitter においてリアルタイムな情報共有がユーザ動詞で行われ，被害状況や安全確認の情報共有がされていた．宮部ら [6] らは大規模災害時の Twitter 上でのコミュニケーションが重要な情報インフラの役割を果たしていたと考え，利用分析から今後の大規模災害での適切な情報提供を実施するシステム開発について知見を持っている．この研究では Twitter のある同じポストを再度自分のアカウントから発信するリツイートや特定のアカウントに対し，情報発信を行うリプライなどの利用状況から分析を行っている．

また，目黒ら [7] はより情報共有を進めるために Twitter のコーパスを用いて，投稿された質問に対し，質問の中からキーワードを抽出し，それの返答となるリプライをコーパスから抽出する研究を実施した．この研究ではポストに含まれる語の抽象化や特徴量を抽出している．また，Twitter は 140 字制限というショートテキストであり，テキストに含まれる情報量が少なくなる．さらに，従来の対話データとは違い，知り合い同士がフランクに対話を行い，表現，話題などが多岐にわたる．そのため，シソーラスと N-gram を用い，特徴量を保持することを提案している．

本研究では特徴量として一部の検索語を特徴語として設定している．特徴語の設定には形態素解析において 2 種類の辞書を使用し，その差異を特徴語としている．

3. コーパス分析

本研究で用いられているコーパスはマイクロブログサービス Twitter に 2014 年 1 月 1 日から 2014 年 12 月 31 日に日本語で投稿されたショートテキストのうち 500,771 件の対話データをコーパスとしている．それぞれの対話データは元の投稿である POST とそれに対するリプライとなる CMNT の 2 つのショートテキストによって構成されている．また，検索クエリに関しても同様に Twitter に投稿された 200 件のショートテキストである．

表 2 にコーパスとして使用された対話データの例を示す．

3.1. 特徴語

POST や CMNT に含まれる語の中には明らかにショートテキストの特徴を示すことができる語や固有名詞などの他の語に比べて出現頻度が低く，特徴的な語として考えられるものがある．本研究ではそれらを特徴語と設定し，検索クエリとなるショートテキストから特徴語の抽出を実施している．

表 3 に特徴語を含むショートテキストの例を示す．

表 2: コーパス（抜粋）

種別	内容
POST	やっとテスト終わった !!! 解放感 !!!
CMNT	お疲れ様っす !!
POST	コンタクト外した瞬間目が痛い…なんだろ
CMNT	コンタクトなんだ !!
POST	帰宅ー！今月はあんまり買い物しないようにしないと
CMNT	おかえりなさーい！

表 3: 特徴語を含むショートテキスト例

内容	特徴語
艦これ 5 話よかった。 瑞鶴かわいい で、瑞鳳は？····	艦これ，瑞鶴 瑞鳳
そういや g のレコンギスタはいつ放送なんですかねえ	g のレコンギスタ

表 3 において “艦これ”¶ という固有名詞やそれに登場する “瑞鳳”，“瑞鶴”などがショートテキストの特徴を示す語として大きな役割を持っていることがわかる．

3.2. ショートテキストのテキスト長

Twitter では投稿するショートテキストに文字制限が設定されており，その制限は 140 字（全角半角は問わない）となっている．また，今回使用したコーパスの平均テキスト長を求めると表 4 のようになった．

表 4: POST と CMNT の平均テキスト長

	平均テキスト長
POST	61.98 文字
CMNT	41.46 文字

3.3. 特定のショートテキスト

ショートテキストには特定のやりとりが POST と CMNT により行われているものが見られた．表 5 にその代表例を示す．

挨拶などで使用される感動詞がショートテキストが POST に登場する場合は期待される CMNT として同様に感動詞が使用されていることが考えられる．また，Twitter のサービスの 1 つであり，他者の投稿を受動的に得ることができるサービス “フォロー”についても同様に “よろしくお願いします”という挨拶を示す語が見られた．

¶艦隊これくしょん-艦これ-の略称．ブラウザゲームとして DMM.com 社が配信しているシミュレーションゲーム．

表 5: 特定のやりとり

種別	テキスト	内容
挨拶	POST	みんなおはよう !! 今日も一日がんばろうね !!
	CMNT	おはようございます。 週末は暖かくなりそうですね。
フォロー	POST	フォローありがとうございます！
	CMNT	よろしくお願いします！

4. 提案手法

本研究では返信元となる検索クエリと類似するショートテキストを POST コーパスへ検索を行うことにより導出し，導出された POST に対応する CMNT を結果として引用するものである．

また，**3** で述べたようなコーパスの特徴を検索手法に取り入れている．

4.1. 検索クエリ

本研究で使用する検索クエリはコーパスと同様に Twitter から入手した日本語で記述された 200 件のショートテキストである．

4.2. 手順

提案手法の手順は以下のようになる．

1. 検索クエリから単語を抽出し，検索語を作成する．
2. 検索語に条件に応じた重みを付与する．
3. TF-IDF における検索を POST コーパスに実施する．
4. 検索結果からテキスト長に応じたスコアリングを実施する．
5. 検索結果の上位のものに付与されている CMNT を出力する．

4.3. 検索語の生成

検索語の生成は検索クエリを形態素解析ソフト mecab[8] によって語分類がされる．また，mecab で使用する辞書は mecab-ipadic-NEologd[9]（以下，NEologd）を使用している．NEologd は複数の Web 上の言語資源から得た新語を追加した辞書である．これにより，比較的新しいサービスである Twitter などの形態素解析への利用に適しているものである．

本研究で用いる検索語は名詞，形容詞，感動詞の 3 つの品詞となる．また，それぞれ変化がある品詞は終止形の形で検索語に抽出される．

また，いくつかの動詞とその活用形に対しても検索語の抽出を行っている．表 6 に抽出例を示す．

表 6 に挙げられる語はショートテキストの中でユーザの意見を示すものとして抽出を行っているものである．

表 7 に検索語抽出例を示す．

表 7 では**洗いたい**という願望を意味する助動詞に続く動詞が表れているので検索語として登録されている．

表 6: 動詞に関する検索語

検索語パターン	説明
動詞 ＋ 助動詞（たい）	願望を示す検索語
動詞 ＋ 助動詞（ない）	否定を示す検索語

表 7: 検索語抽出例

クエリ	検索語
その間の記憶が本当にない	その間，記憶ない
シャワー浴びて顔洗いたい	シャワー，顔洗いたい

4.4. 特徴語の設定

本研究では特徴語の設定を検索に取り入れている．特徴語の抽出には形態素解析ソフトである mecab で使用する辞書を 2 種類使用することで抽出を行っている．使用した辞書を表 8 に示す．

表 8: 使用した辞書

辞書	備考
ipadic	標準的な辞書として使用
NEologd	拡張辞書として使用

検索クエリから検索語の抽出には NEologd を使用しているのだが，同時に ipadic でも形態素解析を行い，2 種類の解析結果を得る．本研究ではこの 2 種類の解析結果の差分を特徴語として採用している．

例として “ISPJ” という語に対し，ipadic は未知語として処理されているのに対し，NEologd では “IPSJ” という名詞として処理されている．また，“情報処理学会” という語に対しても ipadic では “情報処理” と “学会” の 2 つの名詞で処理されているのに対し，NEologd では “情報処理学会” の 1 語の名詞として処理されている．

この場合，“IPSJ”，“情報処理学会” と 2 語は特徴語として検索語に登録される．

4.5. 検索語の重み付与

検索語には重みを付与し，その語の持つ重要性を検索結果に反映している．重みは検索語に応じた 6 種類のものを設定した．表 9 に設定された重みと適用される条件について示す．

感動詞にはほとんどの挨拶が含まれており，感動詞の重みを優位にすることによって多くの対応する挨拶を検索によって導出することができる．

また，処理の経過においてひらがな 1 文字の語が名詞として検索語に登録してしまう可能性がある．そのため，ひらがな 1 文字の検索語はノイズとして処理し，検索結果には関与しないようにした．

表 9: 重みと適用条件

重み	適用条件	備考
0	ひらがな 1 文字の検索語	ノイズとして処理
2	願望を示す検索語	ユーザの意見の反映
2	否定を示す検索語	ユーザの意見の反映
3	感動詞	特定のやり取りを多く含むため
4	特徴語としての検索語	特徴語として評価
1	その他の検索語	

4.6. テキスト長に関するスコアリング

3.2 で述べたように，POST と CMNT それぞれの平均テキスト長にはおよそ 20 文字の差があることがわかった．このことから，それぞれの POST には POST のテキスト長に応じて期待される CMNT の適したテキスト長があると考えられる．

そのため，それぞれの POST の文字数に対して対応する CMNT の文字数を調べた．そして，その結果をもとにそれぞれの CMNT テキスト長の出現確率を算出した．また，その出現確率を使用し，検索結果から得られたランキングを式（1）で再スコアリング及び再ソートを行った．

$$S' = S + \log_2(P \times 100) \times \alpha \quad (1)$$

S は検索結果から得られたスコアを示し，P はテキスト長の出現確率を示す．α はテキスト長スコアリングに関する重み設定であり，今回の実験では 1/100 とした．S' はテキスト長スコアリング適用後のスコアを示す．また，期待されるテキスト長から大きく外れる場合はマイナス評価が与えられるようになっている．

5. 実験結果

本研究において 2 種類の RUN を作成した．表 10 に作成した RUN を示す．

表 10: RUN

RUN	説明
BASE-RUN	特徴語抽出などをおこなったもの
LENG-RUN	BASE-RUN に加え，テキスト長スコアリングを行ったもの

また，それぞれの RUN において STC の公式評価結果の Mean nERR@5 を表 11 に示す．

本研究において顕著にその効果が現れたのは検索クエリから特徴語の抽出が効果的に出来た場合である．例えば，

表 11: MAP

RUN	Mean nERR@5
BASE-RUN	0.3620
LENG-RUN	0.3825

「これだからゆとり世代は…」みたいなことを言ってる人の方が常識ないことも多々あるしね。

という検索クエリに対しては “ゆとり世代” という語を特徴語として抽出し，得られた類似する POST も

これだから、ゆとり世代は、、、という言葉ほどむかつく言葉はない。世代ってのは、君たち先輩が作ったものであり、私達ゆとり世代は作られた側である。それを、まさか自分達は悪くない！この世代のせいだとでも思っていますか？しかも、世代世代って！個人個人で見れない時点で、器がしれてる。

と十分に類似性のあるものが得られた．また，それに対応する CMNT も

たまたま「ゆとり世代」って言われてるだけで、いつの時代もこれだから最近の若い奴らは、、、って言ってる気がする

と検索クエリに対する結果として十分な内容と考えられる．

また，その他に効果的だったものとして挨拶などの感動詞を含む検索クエリも同様の挨拶を含む POST が導出でき，その結果としても十分な CMNT を引用できていた．

また，テキスト長スコアリングにおいても表 11 が示すようにスコアが向上しているため有用な手法であることが分かる．

6. 考察

検索クエリから特徴語を効果的に抽出出来た検索クエリに関しては十分な効果が期待できることがわかった．また，挨拶などを含む感動詞の優遇やユーザに主張の反映としての希望動詞・否定動詞の検索語化なども効果的であることがわかった．また，テキスト長に関するスコアリングを実施することで類似する POST のみを評価するだけではなく，期待されるテキスト長よりも極端に短く意味を持たなくなってしまう CMNT のマイナス評価などが可能となり，CMNT の妥当性も検証することができるようになった．

しかし，検索クエリによっては特徴語が抽出できない場合や検索語の品詞を限定しているため，検索語が全く抽出できない場合も見られた．また，CMNT の評価は現在テキスト長のみでしか評価されず，十分な評価を行うためには POST と CMNT に登場する語の全く新しい手法を考案する必要がある．

また，今後の展望として特徴語の持つ特徴量の抽出が考えられる．本研究での特徴語の抽出は検索クエリのみでしか行っていない．抽出の範囲をコーパスにも行うことによってより多くの特徴語を入手することができ，その出現頻度などから特徴語の特徴量を設定し，その数値を検索のスコアに反映できることが考えられる．

7. おわりに

Twitter や Weibo といったマイクロブログはユーザ数が多く，また，その更新頻度も他のブログサービスに比べて多くなる．さらにショートテキストでやり取りされるマイクロブログではより会話に近い対話のやり取りがされている．そのため，今回使用した NEologd と ipadic の差分を用いた検索手法は有用であると考えられる．

また，マイクロブログ上での情報共有は企業にてっても重要なインフラなっており，平時の宣伝だけではなく株式会社 LINE は 2016 年 4 月 14 日に発生した熊本地震の際には Twitter などで自身のサービスの一部無料化などを告知している．

このようなことからも，自動リプライシステムによる情報共有は重要であると考えられる．

以上

参考文献

[1] Twitter，http://twitter.com/（2016 年 4 月 14 日 参照）

[2] Weibo，http://weibo-japan.blog.jp/（2016 年 4 月 14 日 参照）

[3] NTCIR，http://research.nii.ac.jp/ntcir/（2016 年 4 月 14 日 参照）

[4] NTCIR-12，http://research.nii.ac.jp/ntcir/ntcir-12/（2016 年 4 月 14 日 参照）

[5] STC，http://ntcir12.noahlab.com.hk/japanese/stc-jpn.htm（2016 年 4 月 14 日 参照）

[6] 宮部真衣, 荒牧英治, 三浦麻子,“東日本大震災における Twitter の利用傾向の分析 ”. 情報処理学会研究報告,2011-GN-81,No.17,pp.1-7（2011）

[7] 目黒豊美, 東中竜一郎, 杉山弘晃, 南泰浩,“意味属性パターンを用いたマイクロブログ中の発言に対する自動対話行為付与 ”. 情報処理学会研究報告,2013-SLP-98,No1,pp.1-6（2013）

[8] MeCab，http://taku910.github.io/mecab/（2016 年 4 月 14 日 参照）

[9] mecab-ipadic-NEologd : Neologism dictionary for MeCab，https://github.com/neologd/mecab-ipadic-neologd（2016 年 4 月 14 日 参照）

謝辞

本研究における実装並びに評価に際し，Twitter より提供されたデータセットを利用しています．また，評価型ワークショップ NTCIR-12 Short Text Conversation(STC) への参加と遂行に関して General，Program 並びに Task オーガナイザである宮尾 祐介氏並びに東中 竜一郎氏にご助力いただきました．ここに記して謝意を表します．