自動リプライのためのマイクロブログへの効率的な属性付け

森下　雄太†　　渋川　翔太††　　佐藤　隆士†††

† 大阪教育大学大学院 教育学研究科　〒 582-8582 大阪府柏原市 旭ケ丘 4-698-1
†† 大阪教育大学 教育学部　〒 582-8582 大阪府柏原市 旭ケ丘 4-698-1
††† 大阪教育大学 情報処理センター　〒 582-8582 大阪府柏原市 旭ケ丘 4-698-1
E-mail: †morishita@ss.osaka-kyoiku.ac.jp , ††shibu@ss.osaka-kyoiku.ac.jp, †††sato@cc.osaka-kyoiku.ac.jp

あらまし　Twitter や Weibo などのマイクロブログサービスでは様々なポストととそれに対するコメントでユーザ同士のコミュニケーションが行われている．本研究はそのコメントの自動化を試みるものである．コメントは自動生成をするわけではなく，過去に投稿された類似のポストを検索によって抽出し，それに対応するコメントを提案するものになる．また検索の精度向上と効率化に向けてコメントへの属性付与を行っている．また，その他に未知語・固有名詞の抽出や語数による重みづけなどの工夫も行っている．

キーワード　マイクロブログ，SNS，自動リプライ，マイクロブログ，ショートテキスト，コーパス拡張

1. はじめに

マイクロブログサービスとは Twitter [1] や Weibo [2] を代表する比較的短いテキストをやり取りし，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．それぞれのサービスではユーザの考えなどが投稿され，それに対して別のユーザが返答を行うコメントサービスが実装されている．そのコメントには投稿者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コメントを自動的に作成することを目標とした研究となる．コメントは自動生成するわけではなく，過去の投稿のうちポストとコメントがペアとなったコーパスから検索を行い，適切であると思われるコメントをユーザのポストへのコメントとして提示するものである．

評価型ワークショップ NTCIR-12 Short Text Conversation [3] [4]（以下，STC）において提示されたコメントは表 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 Weibo に投稿された 5,648,128 ペアをコーパスとして検索を行った．

検索の手法としては投稿に内容に準じた属性を付与することによって検索効率を上げる属性検索と未知語や固有名詞の抽出を行いクエリしていることで検索精度の向上を目指した．

コーパスの特徴として，もともと中国語で記入されたものを英語に機械翻訳されたものを使用していることと，同一のポストやコメントが多く見られることが特徴として挙げられる．今回の研究においてもこれらの特徴を検索手法の工夫として取り入れている．

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，提案手法として検索語の作成から未知語・固有名詞の抽出，語数に関するスコア倍率，コメントに対する属性付けなどの検索手法について述べる．**5** では検索の結果についての提示と考察を行う．**6** では今後の課題について述べ，**7** でまとめを行う．

2. 関連研究

マイクロブログサービスや携帯端末の普及により個人が発信するテキストデータは膨大なものとなっている．そのため，テキストに属性を付与する研究には様々なものがある．

乾ら [5] の研究もそのひとつであり，個人の記述する意見を評価，要望，不安，認識，賛否の 5 つに分類を行い，特に評価に関するものを研究対象としている．さらに，テキストの長さに応じて語句，文，文章と 3 つのレベルに分類し，語句の共起情報などを用いて実験を行っている．しかし，本研究で扱うテキストは総じて短いものとなっており，共起情報などではスコアリングできないものも多い．

また，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．本研究と目黒らの研究との違いは未知語・固有名詞に着目している点にある．目黒らの研究では形態素解析を用いているが，未知語・固有名詞に関しては N-gram を使用し，活用を行っていが，本研究では未知語・固有名詞をテキストに含まれる最大の特徴量としている．さらに，フランクな言い回しが多いマイクロブログサービスにおいて感情語の活用の有無も目黒らの研究と異なる点となる．

3. コーパス分析

本研究で用いられたコーパスは中国のマイクロブログサービスである Weibo に投稿されたショートテキストのポストとコメントの 5,648,128 ペアである．また，検索クエリとして使用するポストクエリも同様に Weibo に投稿された 100 件のポストとなる．表 2 にコーパスとして使用したペアの例を示す．

表 2　コーパス (抜粋)

種別	内容
POST	42 children in Xinjiang Kashi skin Li Village, every day going to school road which must be taken.
CMNT	Too dangerous
POST	You think yourself?
CMNT	With is not well good, is own point of view decision, if content, that is good
POST	This boy is lazy enough, alit from a plane, but must sit on the suitcase makes others drag away.
CMNT	Real master

3.1 同一の投稿

コーパスの中には投稿の内容が別の投稿の内容と完全に一致している投稿が見られた．表 3 にポストのものを，表 4 にコメントのものの TOP10 件をそれぞれ示す．

表 4 を確認すると比較的短い文章が多くコメントとして使用されていることがわかる．

また，表 4 中には同一内容のものが別の語として扱われいるが，これは中国語の元データを英語翻訳した際に同一内容のものになってしまった影響である．

3.2 未知語・固有名詞

コーパスは Weibo のものを利用しているため，もともと中国語で記入されていたものを機械翻訳によって英語に翻訳したものである．そのため，コーパス中には中国における人名や地名，その他の固有名詞が多く残っている．また同様に未知語に関してもショートテキスト内に残っている．

4. 提 案 手 法

本稿の提案手法は未知語・固有名詞抽出における適性文書の抽出と文書に付与された属性による属性スコアリングである．

表 3　POST に含まれる同一内容投稿数（TOP10）

投稿数	内容
2746	Friends good morning, today's air quality.
2318	Friends good morning, today air quality.
1102	We first senile is always not the appearance, but is that reckless rushing vigor.
955	Goes ahead, is the life most gorgeous luxury, is the brightest freedom.
946	Travels the biggest advantage, cannot see many people, sees many beautiful sceneries, but walks is walking, under a spell of good or bad fortune, knew itself suddenly.
862	Future you, will certainly thank go all out you now diligently.
859	Today air quality of Beijing.
848	In the life at least must have two impulsions, one time is the love of acting boldly regardless of one's safety, one time is the travel of going ahead.
798	In life 10 helpless matters. . .
784	The people have two roads to walk, must walk, wants to walk, you must walk attractively the road that must take, may take the road that wants to take.

表 4　CMNT に含まれる同一内容投稿数（TOP10）

投稿数	内容
14,830	Ha
8,096	Ha
8,075	Ha-Ha
4,443	ng
4,297	Ha
4,222	Good
3,676	Good
3,468	Yes
3,216	Like
3,209	Great

4.1 手　　順

手順は以下のような順序で行う．

（1） ポストクエリから検索クエリの抽出を行う

（2） ポストクエリから未知語・固有名詞の抽出を行い，検索クエリに追加する

（3） ポストクエリ，コーパスの文書それぞれに属性付けを行う

（4） コメントの語数に応じた重み付けを行う

（5） 作成された検索クエリからコーパスに対して検索を行う

（6） 検索結果に対して属性によるスコアの補正を行い再ソートする

4.2 未知語・固有名詞抽出

4.2.1 単語の切り出し

本研究では単語ごとの検索を行っている．そのため，TreeTagger [8] を用いて形態素解析を行い，名詞，動詞，形容詞，副詞を検索語の候補として抽出を行った．

表 5　検索語の抽出例

テキスト	検索語
Arrives at Qingdao, compared with an impression better city. To has remembered the train ticket to forget to give money	Arrive, Qingdao, compare, impression, better, city, has remember, train, ticket, forget, give, money
Is far away from the city, peacefully.	far, away, city, peacefully

冠詞や接続詞などの他の品詞に関しては文章的に意味を持たないため今回は除外して実験を行った．また，その他にも do not の do や完了形の文法に用いられる have なども除外している．さらに，複数回出現する語に関しても 2 回目以降の語は検索語に含まないようにした．

4.2.2　未知語・固有名詞抽出

検索精度向上のために未知語・固有名詞に対しても重要な検索語として検索を実施している．

未知語・固有名詞は **4.2.1** で実施した単語の切り出しにおいて TreeTagger により固有名詞として出力されたものに加え，未知語として判断されたものの一部も検索語に加えた．

また，連語などの明らかに 2 語以上の語を使用して表現されている語に関しても固有名詞として検索語とした．

この操作によって挨拶や地名，行事名，人物名などの抽出を行うことができた．

表 6 に未知語・固有名詞の抽出例を示す．

表 6　未知語・固有名詞の抽出例

テキスト	抽出された語
Arrives at Qingdao, compared with an impression better city. To has remembered the train ticket to forget to give money	Qingdao
Lantern Festival supper. Draws a charge everybody Lantern Festivalto be joyful I in	Lantern Festival
Happy New Year, who discussed that PPT I kick whose fart fart	Happy New Year, PPT

4.2.3　検索語のオプション

検索語にはオプションを追加することができる．オプションを表 7 に示す．

表 7　オプション

オプション	説明
weibo;;1	weibo を検索語とする
weibo;;2	weibo を重み 2 の検索語とする
weibo;;1;;e	weibo を必須語とする
weibo;;1;;n	weibo を除外語とする
weibo;;1 twitter;;1;;p	weibo もしくは twitter を 1 つの検索語とする

これらのオプションは必要に応じて付与され，特に出現頻度の低いと思われる語に関して必須語オプションや高い重みを設定することで検索精度の向上が期待できる．

4.2.4　拡　　張

検索語にはいくつかのクエリ拡張を施している．表 8 に拡張を示す．

表 8　クエリ拡張

種別	拡張対象	拡張後検索語
否定動詞文法	not + 動詞	not 動詞
語尾変化拡張	語尾変化された語	arrived;;1
		arrive;;1;;p
連語拡張	連語	HAPPY NEW YEAR;;1

否定を意味する not は除外するべき語ではあるが文意的に重要な語となるため否定の対象となる語の頭に付与し一つの検索語として追加される．

また，語尾変化された語に関しては原型を同時にパラレル検索を実施するように検索語に追加される．さらに **4.2.2** で述べた挨拶や固有名詞などは複数の語で構成されることが多く，それらも連語として一つの検索語となる．

4.3　語数に関するスコアリング倍率の設定

表 3,4 にある通り，コーパスの中には同一内容の文書があり，それらを出現頻度でソートした結果，上位になるものはかなり小さな語数で構成されていることがわかる．

本研究ではこの語数についても着目し，短い語数がスコアが向上する工夫を行っている．

しかし，語数が小さいほどテキストとしての情報量は小さくなってしまう．そのため基本語数は 3 とした．また，3 より小さい語数は 3 として扱う．

式 1 に語数に関するスコア倍率の算出方法を示す．

$$W_n = \frac{1}{\sqrt{\frac{n}{3}}} \tag{1}$$

W_n:語数におけるスコア倍率

n:語数

また，それぞれの語数における倍率は表 9 のようになる．

表 9　語数によるスコアリング倍率

語数	スコア倍率
1-3	1
4	0.866
5	0.775
10	0.548
20	0.387

4.4　属性スコアリング

4.4.1　属 性 付 け

本研究ではコーパスと検索クエリそれぞれに属性付けを行った．

コーパス内の属性付けは **3.1** で述べたように同一の内容がコーパス内に多く含まれていることを考慮し，出現頻度が高いものとそれに類似するものを優先的に属性付けを行った．一つのコメントには複数の属性が付与されることもある．

類似する投稿の抽出は基本となる投稿からキーワードと除外される語を設定し，コーパス内から抽出を行った．また，付与した属性は以下の表 10 に示す．

表 10　属 性 一 覧

属性	説明	付与数	付与コメント例
positive	ポジティブなコメント	150,420	Attractive
agree	同意するコメント	141,373	Right
laugh	笑い声を示すコメント	45,478	Ha
surprise	驚きを表現するコメント	14,959	Oh
beautiful	美しさを示すコメント	14,468	Really beautiful
praise	賞賛するコメント	14,326	Has the talent
lovable	かわいさを示すコメント	6,726	Lovable
cheer	元気づけるコメント	6,100	Come on!
greeting	挨拶のコメント	4,923	Good night

属性付けを行ったことにより，398,776 件のコメントに対して属性を付与した．これはコメント全体のおよそ 7 パーセントにあたる数値となる．

また同様の手法によりポストクエリの 100 件にも属性付けを行った．

4.4.2　属性付けによるスコアリング

属性付けによるスコアリングは 4.2 で算出されたスコアに属性ごとに定められたスコアを加算するものとなる．

加算対象となる文書はポストクエリに付与された属性と一致する属性を持つ文書である．イメージを図 1 に示す．

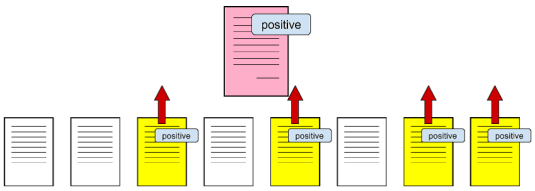

図 1　属性スコアリングのイメージ

また，属性のスコアはそれぞれに設定されている．表 10 から付与数には大きな差があることがわかる．そのため，付与数が少なく出現確率が小さくなる属性には大きなスコアを設定している．

属性ごとのスコアは以下の式 2 で求めている．

$$S_n = \sqrt{\log \frac{\sum a_k}{a_n}} \times 0.1 \qquad (2)$$

S_n:属性ごとのスコア

a_n:属性数

また，それぞれのスコアは表 11 のようになる．

5.　結　　果

本研究では複数の RUN を作成した．その中で 3 種類の RUN について検証を行う．作成した RUN を以下の表 12 に示す．

作成された 3 種類の RUN において Useful もしくは Coherent であるもの適合文書とし，それぞれの MAP を算出した．表 13 に示す．

表 11　属性ごとのスコア

属性	スコア
positive	0.1089
agree	0.1106
laugh	0.1330
surprise	0.1475
beautiful	0.1479
praise	0.1480
lovable	0.1557
cheer	0.1567
greeting	0.1587

表 12　作成した RUN

RUN	語検索	語数	未知語・固有名詞	属性
BASE RUN	○	○	×	×
EXPAND RUN	○	○	○	×
ATTR RUN	○	○	○	○

表 13　作成した RUN

RUN	MAP
BASE RUN	0.4194
EXPAND RUN	0.5477
ATTR RUN	0.4595

表 13 より，EXPAND RUN は BASE RUN に比べて 30 パーセントのスコア向上が見られた．しかし，ATTR RUN については EXPAND RUN に比べて 17 パーセントの数値の低下が見られた．

よって，未知語・固有名詞抽出は今回のタスクにおいて有用であることが示せた．

EXPAND RUN がスコアを減少させることになってしまった原因として付与した属性の数と種類にあると考えられる．

本研究では属性は表 4 が示す同一投稿が多いコメントを優先し，属性付けを手動で行った．そのため，属性の方向性に偏りが発生してしまった．表 10 にはポジティブ方向での属性が大半であり，ネガティブ方向のコメントはない．そのため，ネガティブ方向のコメントを求めるポストクエリに対応できなかったことが考えられる．

6.　今後の展開

本研究において，属性付与数が小さかったことが ATTR RUN の数値を低くしてしまったことが考えられる．

本研究では属性付付与は付与する範囲については手動で行っている．これは大変非効率であるため，この点を効率化するために再帰的な属性付与を考案している．

再帰的な属性付与は属性付与を行っていない状態のポストクエリでの検索結果（EXPAND RUN 相当）からコメントに付与されている属性を参照し，ポストクエリに属性を自動的に付与するというものである．ポストクエリへの自動属性付与イメー

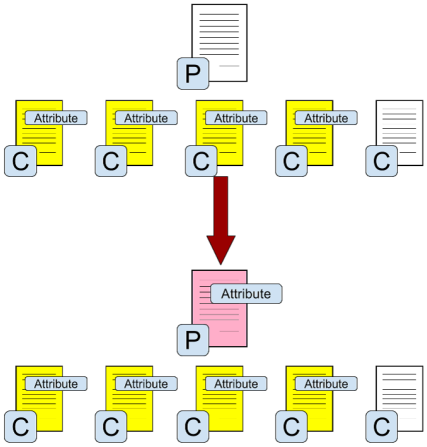

図 2　ポストクエリへの自動属性付与イメージ

ジを図 2 に示す．

これを実施することによって手動で行っていたポストクエリへの属性付与を自動化することができる．

また，一定以上のポストに属性付与を実施することによって新しくコメントへの属性付与も自動化も期待できる．

同一属性が付与されたポストグループを検索クエリとし，検索結果のコメントの中でスコアが高く，多くのポストクエリから支持されたコメントに属性を付与する．コメントへの自動属性付与イメージを図 3 に示す．

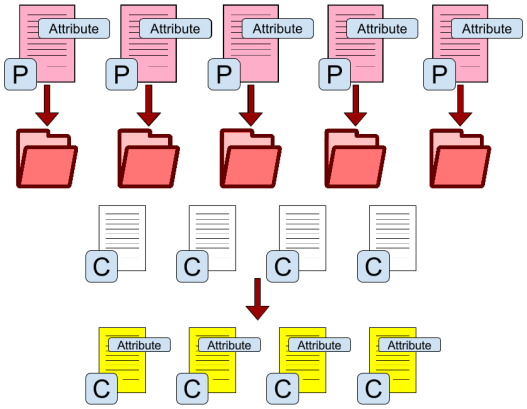

図 3　コメントへの自動属性付与イメージ

また，今後の展望として日本語における自動リプライのシステムとして Twitter を用いた研究が大阪教育大学で進められている．コーパス分析の結果，英語で実施された本研究とは違い，語数による有意差は見られないものの未知語・固有名詞の抽出として 2 種類の形態素解析器を用いてその差分を利用する手法がとられている．

7. おわりに

本研究では未知語・固有名詞抽出による自動コメントシステムは一定の効果を示すことができたが，属性スコアリングによるシステムはシステムを阻害してしまう形となった．

しかし，属性スコアリングも属性の種類や属性付与された投稿を増加させることにより効果を示すことが期待できる．

STC のシステムの精度をより向上させることができれば，マイクロブログサービスにおけるリプライだけではなく，機械との会話やユーザが求めている情報の提供など，様々な分野に応用できる技術だろう．

文　献

[1] Twitter
`http://twitter.com/` (2016/2/12)
[2] Weibo Japan
`http://weibo-japan.blog.jp/` (2016/1/7)
[3] NTCIR
`http://research.nii.ac.jp/ntcir/` (2016/1/7)
[4] Short Text Conversation at NTCIR
`http://ntcir12.noahlab.com.hk/stc.htm` (2016/1/7)
[5] 乾孝司, 奥村学,”テキストを対象とした評価情報の分析に関する研究動向”. 自然言語処理 vol13,Num3,pp.201-241（2011）
[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] TreeTagger - a language independent part-of-speech tagger
`http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/` (2016/1/7)

謝　辞

本研究における実装並びに評価に際し，Weibo 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 Lifeng Shang 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．