なにをつぶやいているのか?: マイクロブログの機能的分類の試み

菊井玄一郎
岡山県立大学　情報工学部

kikui@cse.oka-pu.ac.jp

1　はじめに

マイクロブログは数年前より爆発的な普及を見せており、代表的なサービスであるtwitterはユニークユーザ数が1億人以上，一日2億5千万件以上のメッセージが投稿されている[1].

マイクロブログは内容的にも多様性に富んでおり，田中ら [9] が指摘している通り，特定のコミュニティでのみ意味のある内容から，広く一般に有用な内容まで，また，リアルタイム性の高いものから低いもの，ディスカッションの一部を構成するものから何の脈絡もなく突然書き込まれるものまで様々なメッセージがある．これはマイクロブログ（以下の議論ではtwitter[2]を想定）が単に「長さが (140 文字に) 制限されたブログ」というだけではなく，発信者を指定してメッセージを購読したり，他のメッセージを引用・応答したりできるコミュニティ形成支援機能を提供していること，携帯機器上で簡単かつリアルタイムに投稿・閲覧できるソフトウェアが普及していることなどによって，コミュニティサイト的な使い方やBBS，チャット的使い方など様々な使い方ができることによる．また，twitterでは身の回りのごく私的な出来事や自身の心身の状況を断片的かつリアルタイムに発信することが許容されており，これがコミュニティにおける一種の自己表現として利用されていることも特に「特定のコミュニテイでのみ意味を持つ」と考えられる情報が発信される一因である．

既存研究の多くは，一般的なトピック（一般に「ニュース性」のあるもの）に関するメッセージと，それ以外の私的なメッセージとに分け，一般に後者はノイズとされている（例えば [2]）．しかしながら，身の回りの出来事や状況を発信する「プライベートなメッセージ」についても，情報源として有用であり，たとえば，時間情報や地理情報と統合することにより，地震の発生やインフルエンザの流行など実世界のイベントが検出できる [3] [1]。

「広く一般的に有用なトピック」に関するメッセージについてはトピック分類や内容に基づくクラスタリングなどによる細分類が有用であり，実際，そのような研究が行われている ([7],[8] など)．一方，「プライベートなメッセージ」について，その内容の分布や経時的な変化はどうなっているかなどは明らかになっているとは言いがたい。

そこで，本稿では第一歩として、「プライベートなメッセージ」を主な対象とした分類体系について検討する。特に、マイクロブログに特徴的と言われている、リアルタイム性の高い言明、自身あるいは近い人々の行為や状況、意見や感想の表明などに着目して分類体系の構築を試みる。

2　既存の分類体系

本節では、マイクロブログに対する既存の分類体系について紹介する。これらはいずれも自動分類手法の研究における分類体系として紹介あるいは提案されているものであるが、ここでは分類手法については立ち入らない。

Castilloら [2] は情報の有用性 (interesting-ness) を評価するという観点から，まず，全体を「ニュース性の高い (newsworthy)」もの (NEWS) と「個人的な会話 (private conversation)」(CHAT) とに分け，前者をさらに「信頼できる書き込み」と「それ以外」の2つに分けている。この体系では、たとえば、身の回りの体験などは多くの人に取って「興味の対象外」であるCHATに分類され、それ以上の分類がなされていない。

Sriramら [4] はメッセージをNews, Opinions, Deals, Events, Private Messagesの5つに分類する方法を提

[1] Twitter Inc. March 2010
[2] http://www.tiwtter.com/

案している。この分類は読みたいメッセージを特定する（絞り込む）際の手がかりとなるように設計されており、最後の 1 つを除いて twitter における public なメッセージの代表的な内容をカバーしていると考えられる。しかしながら、Opinions や Events などは private messages と区別が難しく、定義も明らかにされていない。

Zhang ら [5] は tweet を会話における発話と捉え、Searle の発話行為 (speech act) の分類に若干の変更を加えた次の 5 分類を提案している。

Zhang et. al	Searle(1975))
Assertive(陳述)	Assertive
Question(質問)	Directive
Suggestion(提案)	Directive
Comment(コメント)	Expressive
Misc(その他)	Commissive+Declarative

この分類はメッセージをその働きかけの機能によって分類したもので Castillo らによる分類で CHAT とされているものも対象となっている点から興味深い。しかしながら、マイクロブログに特徴的である状況即応的な書き込みは全て陳述に分類され，その詳細が分類できない。また、そもそも、書き込みの前提として「独白（つぶやき）」という聞き手へのコミットが非常に低い状況において，Searle の言う発話行為の遂行条件（たとえば、正常入出力条件や誠実性条件など[3]）が成り立っているかどうかは疑問であり，本来の意味での発話行為タイプの分類と考えるのは無理があると思われる．但し，発話行為という視点で発話を分類することは著者らの主張する通り投稿者の行動パターンを理解したり，メッセージを検索したりする上で意味があると考えられる．

田中ら [9] は tweet の検索を支援することを目的として、「情報発信性」「リアルタイム性」「社会性」「有用性」の 4 つの分類軸を提案している。このうち「情報発信性」は情報を意図的に発しているかどうか、「有用性」は既にあげた分類と大きく異なるのは、多次元的な分類になっていることである。他の分類体系との関連で言うと、情報発信性、および、有用性、社会性が低いものが CHAT あるいは private messages に対応すると思われる

以上をまとめると次のようになる。

- ほとんどの分類体系において、私的なメッセージ (private messages) とそれ以外のニュースや意見などとを分け、前者については無価値とみなす，あるいは，さらなる分析・分類を行なっていない．
- 多くが一般的なメッセージとして News 的なものを想定している．
- 「私的なメッセージ」を視野に入れた分類としては発話行為タイプによるものがある

そこで，本研究では特に読み手への働きかけがあるメッセージについては Zhang ら [5] の発話行為の分類を用いることにした．

3 分類体系の構築

3.1 分類の単位

従来研究は 1 つのメッセージ (tweet) に対して 1 つの分類カテゴリを割り当てている．しかしながら、1 つのメッセージにはしばしば複数の文（節）が含まれ、それらが異なった機能（カテゴリ）を持つとき，1 つの機能に絞るのは難しい．たとえば、次の例で第一文は「行動予定」，第二文は「挨拶」であるが，たとえば，全体を「挨拶」とするのは無理がある．

「寝よう．おやすみなさーい.」

そこで，ここではカテゴリ付与の単位を「文」とした．文の境界（文末）は句点を基本としたが，句点がなくても文法的に文末である場合はそこを境界とした．

3.2 基本的な考え方

メッセージ中の各文を次のような特徴によって分類することにした．

1. 発話行為タイプ : 発話内行為に限定して，陳述型のものと，それ以外の依頼，挨拶など対話文に特徴的なものとに大きく分類する
2. 陳述型のものは更に書き手自身に関するもの，外部実世界の記述，抽象的な事柄（真理，法律など）など，書き手との関係，事実性などに基づいて分類する
3. 陳述型以外のものは発話行為の分類を参考に細分類する．

[3] 文献 [6] の用語による

3.3 分類体系

以上の方針に基づいて 1 のような分類体系の試案を作成した.

全体は 9 つの大分類から構成される。以下大分類ごとに概略を説明する。

最初の 3 つの大分類 (A-C) は書き手自身の状態や行動の叙述である。まず A は心的状態や思考に関する記述であり、感情や価値判断、希望、意見などに分かれる。B は身体状態である。基本的には体調と関する記述を想定しているが、身体感覚もここに入れたが両者の境界は微妙である（例 :「空腹だ」)。C は書き手の行動に関する記述であり、過去（完了)、現在（進行)、未来（近接未来、角度の高い予定）に分けられる。

次の大分類 (D) は外部状況に関する記述であり、体験した事実，伝聞が基本的な分類である。自然現象や日付・時刻の 2 つは頻出するので独立したカテゴリーとする。また，予備的なタグ付を行った際に義務的なもの（「空気」のようなもの）が多く見られたのでここに追加した.

大分類 E は真理、規則など事実の記述ではない (non-finite な) 文である。条件（条件節）についてもそれが独立した文の形を取っている限りこの分類に入れた.

F は聞き手への働きかけであり、発話行為と関係が深い。先に述べたように、マイクロブログは特殊なテキスト形態であり、本来の意味での「依頼」などは成立していないことが多いがたとえば要望や願望の一種としての解釈の可能性を考慮して分類を行った。

G 以降は文そのものの分類からやや外れたものである．まず，G はマイクロブログをメモ用紙がわりに使うというものであり、「メモ」で始まるのが特徴である。H は引用のみからなるメッセージ、I はその他分類が困難なものである.

3.4 一致性の予備調査

上記の定義に基づき，200 メッセージ 312 文[4]を対象に 2 名でカテゴリ付与作業を行った．作業結果の一致性を示すカッパ係数は 0.59 であった.

4 おわりに

本稿ではマイクロブログの分類体系について検討した．現在，実際のメッセージテキストについて上記カテゴリの付与を行なっている．一定量のカテゴリ付与を行った後，再度体系を見直して，自動分類による更なる分析を行う予定である.

[4]URL のみの「文」などは除外した

謝辞

分類基準の策定にあたって，当研究室の麦谷智行，原田遼の両君には分類基準へのコメントおよび実データに対する分類作業を実施して頂いた。記して感謝する。

参考文献

[1] E. Aramaki, S Maskawa, and M.Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *EMNLP*, pp. 1568–1576, 2011.

[2] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW Conference*, 2011.

[3] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW Conference*, 2010.

[4] B. Sriram, D. Fuhry, E. Demir, H Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. In *ACM SIGIR*, pp. 841–842, 2010.

[5] R. Zhang, D. Gao, and W. Li. What are tweeters doing: Recognizing speech acts in twitter. In *2011 AAAI Workshop (WS-11-05)*, pp. 86–91, 2011.

[6] 石崎雅人, 伝康晴. 談話と対話 (主に 2 章). 東大出版会, 2001.

[7] 西田京介, 坂野遼平, 藤村考, 星出高秀. データ圧縮による twitter のツイート話題分類. In *DEIM Forum 2011*, pp. A1–6, 2011.

[8] 黒澤, 竹澤. ユーザの返信行動に着目した投稿およびユーザの分類. 言語処理学会第 17 回年次大会, pp. 460–463, 2011.

[9] 田中淳史, 田島敬史. twitter のツイートに関する分類手法の提案. In *DEIM Forum 2010*, pp. A5–4, 2010.

表 1: マイクロブログメッセージの分類体系案

大分類	中分類	表現例	備考
A. 心情	1. 感情	むかつく, うれしー	明確な対象がある場合も含める.
	2. 評価	良い，悪い，値段が安い	いわゆる評価表現を含むもの
	3. 願望	～してほしい，～だったらなあ，～ならいいのに．～したい.	
	4. 意見	～だと思う．～すべき．～したらよい	自分の行動の意志は C
	5. 推測・憶測	～だろう．～かも知れない	
B 身体状態	1. 身体状態	お腹が痛い．腹減った．風邪を引いている．熱がある.	体のコンディション．病気.
	2. 身体感覚	～が冷たい	体調にからまないもの
C 話者の行動・体験	1. 過去	～した．～してしまった．蹴られた	意図的でない場合も含む
	2. 現在	～している．～なう.	話者の所在 (現在位置) も含む
	3. 予定	～しよう．～するぞ，～する	願望は含まない．意志、義務は含む
D 外界の状況	1. 事実・他者の行動		過去も，進行中も含む
	2. 伝聞	～だそうだ．～だって.	推測は A へ
	3. 義務	～しなければならない．期待されている.	
	4. 自然現象	雨が降ってきた．あ，地震． 冷える.（天候が）寒い.	身体感覚表現でも天候に関するものはこちら
	5. 日付・時刻	もう夜だ.	
E 真理等	1. 真理・格言	万物は流転する	
	2. 規則	9 時に出社すること	
	3. 条件	万一～だったらね	反実仮想も含む
F 他者への働きかけ	1. あいさつ	おやすみ.	叫び (例:ああっ) も含む．謝意表現もここ
	2. 名前呼び	～くん.	存在しないものへの呼びかけも含む
	3. 命令・依頼・勧誘	～して．～をお願い.	願望との違いは直接働きかける表現．警告・威嚇もここに入る
	(4. 質問)		保留
	5. 拒否，抗議	なんでよ	
G メモ		「メモ」「自分用メモ」から始まる	自分用メモ
H 引用		”RT”,”QT”で始まる	引用のみのもの
I その他			単語のみ等判断の難しいもの