Twitter 上の未知語の意味推定方式
A method for estimating unknown word meanings on Twitter

牛久保 佑樹† 藤田 茂‡
Ushikubo Yuki Shigeru Fujita

1 はじめに

インターネットの普及によってユーザがニュースサイトやブログなどから情報を得る機会が徐々に増えている．これによってユーザは様々な新情報を手に入れることができるようになったが，未知語の理解がこれら新情報を閲覧する上で必要不可欠である．本研究ではユーザが未知語理解を容易に行うため，未知語の概要を示すシステムを提案する．

服部ら [1] の研究によると，未知語の出現頻度は Web ニュースサイトよりもミニブログなどの CGM(Consumer Generated Media) の方が高いとされ，特に「スポーツ」や「オタク」などエンターテインメント関連の語の未知語率が高いとされる．ミニブログの例として Twitter が挙げられる．Twitter のタイムラインを図 1 に示す．Twitter におけるユーザの発言は未知語や新語などが多く含まれている．未知語に知識の無いユーザはタイムラインの内容を把握することが難しくなり，コミュニケーションが困難になる問題がある．また，エンターテインメント関連の語は新語やスラングを作りやすく，わからない語を見掛けるたびに Web 検索を行い語を理解することはユーザにとって大きな負担となる．このような点から，Twitter のタイムラインにおける未知語の理解支援を行うシステムはコミュニケーション支援に有効であると考えられる．

2 提案システム

本研究では未知語が名詞であると仮定し，ユーザが入力した未知語がどような動作の対象となるか，どのように形容されるか，どのように使われるか，の 3 つをユーザに提示する．未知語にかかる動詞が「食べる」であればユーザは未知語が「食べ物である」と理解でき，形容詞が「おいしい」であれば「食べ物かつおいしいもの」であると判断できる．最後に，未知語が実際にどのように使われているかとして，未知語が含まれている文を例文として提示する．

図 1: Twitter タイムラインの例

結果の出力フォーマット

検索語：(動) 動詞,(形) 形容詞，例文

本提案システムのフローチャートを図 2 に示す．まず，ユーザは知りたいと思った検索語をシステムに入力する．システムは検索語を含むコーパスを取得し，検索語と関係する語と実際どのように使われているかの例文を返す．実際の出力を表 3 に示す．

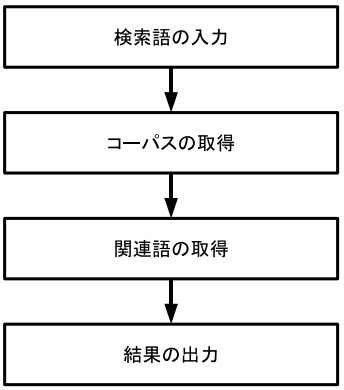

図 2: 提案システムのフローチャート

†千葉工業大学大学院 情報科学研究科 Graduate School of Information and Computer Science, Chiba Institute of Technology
‡千葉工業大学 情報科学部 Faculty of Computer and Information Science, Chiba Institute of Technology

表 1: Sen の動作例

入力：GRAND のロードバイクこうてろ				
G	(G)	記号-アルファベット (0,1,1)	ジー	ジー
R	(R)	記号-アルファベット (1,2,1)	アール	アール
A	(A)	記号-アルファベット (2,3,1)	エイ	エイ
N	(N)	記号-アルファベット (3,4,1)	エヌ	エヌ
D	(D)	記号-アルファベット (4,5,1)	ディー	ディー
の	(の)	助詞-連体化 (5,6,1)	ノ	ノ
ロード	(ロード)	名詞-一般 (6,9,3)	ロード	ロード
バイク	(バイク)	名詞-一般 (9,12,3)	バイク	バイク
こう	(こう)	動詞-自立 (12,14,2)	コウ	コウ
てろ	(てる)	動詞-非自立 (14,16,2)	テロ	テロ

2.1 コーパスの取得

Yahoo!blog 検索 API を用いて検索語を含むブログ記事 10 件を取得する．取得したブログ記事から ExtractContent[2] を用いて本文の抽出を行い，検索語が含まれている範囲をコーパスとして使用する．ExtractContent は html ファイルから本文を抽出するモジュールであり，これを用いることでブログのカテゴリやリンク一覧など本文と関係の無いノイズ部分は除去される．

次に，得られた本文に対して形態素解析器 Sen を用いて形態素解析を行い，本文から単語を抽出する．形態素解析器の辞書には Wikipedia の日本語項目名一覧を固有名詞として追加している．抽出する際，「G/R/A/N/D」「ロード/バイク」など単純な名詞の複合語は 1 つの要素だと考えることが出来る．形態素解析の結果から未知語と判断されたもの，連続するアルファベット，連続する名詞，連続する動詞などは 1 つの要素として抽出する．表 1 の例では「GRAND」「ロードバイク」「こうてろ」が未知語として抽出される．この時，助詞や助動詞は単独では意味を持たないため，これらの要素は除外する．

次に，検索語と抽出した単語の関連性を測るため，両者をおおまかにカテゴリ分けする．分類するカテゴリには Yahoo!ブログカテゴリ中の 20 カテゴリを使用する．使用したカテゴリ一覧を次に示す．

表 2: カテゴリ名

カテゴリ名		
ビジネス	経済	コンピュータ
スポーツ	生活	インターネット
生活	文化	エンターテインメント
科学	趣味	家庭
政治	健康	医学
学校	教育	地域
芸術	人文	

カテゴリ分類には単語と各カテゴリ名とのシンプソン係数を用いる．単語の Web 検索結果数，カテゴリ名の Web 検索結果数，単語+カテゴリ名の Web 検索結果数を調べ，次の式に与える．これをすべての未知語・単語，カテゴリ名との組み合わせに対して行い，値が最も高かったカテゴリ名を未知語のカテゴリとする．シンプソン係数を求める式を次に示す．なお，X は未知語もしくは単語の Web 検索結果数，Y はカテゴリの Web 検索結果数，min(x,y) は値の小さい方を返す関数である．

$$Simpson(X,Y) = \frac{|X| \cap |Y|}{min(|X|,|Y|)} \qquad (1)$$

ここで未知語と同じカテゴリとなった単語を未知語の同類語であると仮定する．一般的に，1 つのブログ記事には 1 つの話題が書かれており，文章中の語はそれぞれ同じような意味を持っている場合が多いと考えられるためである．

次に記事中の 2 つの単語”スマートフォン”，”ギャラクシー S”が同類語であるという例を示す．

> 国内でスマートフォンの価格が下落している．新型の発売に伴い，量販店では旧型のギャラクシー S の価格を 2 万円代にまで引き下げるなど一部機種の大幅な値下げに踏みきっている．中古の人気機種も新型の発売に伴い，3ヶ月前の価格と比べて 4 割から 5 割下落した．消費者が購入しやすい価格帯となり，スマートフォンの普及に弾みがつきそうだ．

このような記事の場合，スマートフォンとギャラクシー S は単語としては別であるが，同様の意味を持つ単語であると考えることができる．

次に，スマートフォンとギャラクシー S のカテゴリがコンピュータであると分類されたとする．「新しいスマートフォンを買った」「ギャラクシー S を使うのは楽しい」という文があった場合，スマートフォンとギャラクシー S は同様の意味を持つ単語であるため，2 つの単語を入れ替えても文章は成立する．

図 3: 処理の流れ

```
1  begin
2  keyword = input(query)  //未知語の入力
3  kDoc = getDocuments(keyword) //未知語を含むドキュメントを取得
4  line[] = extarctLine(kDoc) //ドキュメントから未知語を含む文を抽出
5  word[] = extractWord(line) //文の中に含まれる名詞を抽出
6
7  kCategory = categorize(Keyword); //共起数を用いて未知語を分類
8  for(i=0; i<word.lenght; i++)  //未知語と違うカテゴリの名詞をリストから削除
9          if(kCategory != categorize(word[i]))
10                 word[i].remove()
11
12 for(i=0; i<word.length; i++){
13         wDoc[] = getDocuments(word[i])  //名詞を含むドキュメントを取得
14         line[] = extarctLine(wDoc,word[i]) //ドキュメントから未知語を含む文を抽出
15         verb[] = getVerb(line)   //動詞を抽出
16         adjective[] = getAdjective(line) //形容詞を抽出
17 }
18
19 frequentVerb = getFrequent(verb)  //最頻出語動詞抽出
20 frequentAdjective = getFrequent(adjective) //形容詞を抽出
21
22 eDoc = getDocuments(keyword,frequentVerb,frequentAdjective);
23 example = extract(lDoc,keyword);
24
25 output(keyword+":"+frequentVerb,frequentAdjective,example) //結果出力
26 end
```

未知語や新語はもともとドキュメントに含まれている数が少なく．未知語だけでは充分な量のコーパスを取得できない可能性がある．未知語の他に同類語のコーパスも取得することで，充分な量のコーパスが確保できると考えられる．同類語についても Yahoo!blog 検索 API を用いてブログ記事 10 件を取得し，ExtractContent を用いて本文を抽出する．

2.2 関連語の取得

ここまでで検索語と同類語の 2 種類のコーパスを取得している．それぞれのコーパスから未知語もしくは同類語が含まれている文を抽出し，さらにそこから動詞と形容詞を抽出する．すべてのコーパスから動詞と形容詞を抽出した後，動詞・形容詞の最頻出語を取得し，未知語に関連する動詞・形容詞とする．

例文は未知語を含む文を抽出し，未知語の前後 50 文字，計 100 文字程度を例文として取得する．

図 3 に全体の処理の流れを示す．

3 実験

実験として，タイムライン上に存在する 11 個のキーワードを本手法で実装したシステムに入力し，その出力を得た．実験結果を表 3，実験に使用したブログ記事数を表 4 に示す．

表 4: 実験で使用したブログ記事の数

未知語	ブログ記事数
TINAMI	50
PIXA	100
南鎌倉高校女子自転車部	80
フレアレッド	70
コードマスターズ	60
スケットダンス	50
高速鉄道事故	70
タイバニ	50
ジャローダ	70
HTC ハイロード	90
ピングドラム	70

4 考察

4.1 動詞

動詞について，“高速鉄道事故” や “ピングドラム”（アニメ番組）では “起きる”，“見る” と，動詞としてふさわしいものが抽出できた．しかしながら，“TINAMI” の “頂く” など，動詞を見ただけではどのようなものかわからない動詞を取得する場合があった．ブログは個人の日記のように書かれるため，単語と関連する動詞よりも，感想などを伝える動詞が出現する傾向にある．今回は単純に頻出語を関連動詞として取得したため，未知語と関連性のない動詞を関連動詞として取得してしまったものだと考えられる．係り受け解析を行い，未知語と直接関連のある動詞を取得する，といった解決策が考えら

表 3: 実験結果

TINAMI，動:頂く，形:null, 例:TINAMI もそうなんけど、やはり投稿者の数は少ないなあ
PIXA，動:言う，形:null, 例:TINAMI や PiXA など投稿サイト、新規会員登録が突如 50〜100 倍に（INTERNET Watch）
南鎌倉高校女子自転車部，動:思う，形:null, 例:バガタウェイの新刊がでないか COMIC BLADE の HP 見てみたら、南鎌倉高校女子自転車部って自転車漫画が始まった模様
フレアレッド，動:買う，形:null, 例:フレアレッドのほうが深紅で DSL のが明るい朱色の入った明るい赤
コードマスターズ，動:遊ぶ，形:null, 例:コードマスターズから PS3/XBOX360 用ソフトとして発売される『F1 2011』のゲームプレイトレーラーが公開!
スケットダンス，動:食べる，形:null, 例:スケットダンスを見て、明日からのバイトに備えているほうじちゃでした！
高速鉄道事故，動:起きる，形:null, 例:「あらゆる犠牲に目をつぶって, 表面的な国威のみを追求すること」に大躍進政策の意義があるのだ. だから, 高速鉄道事故で中国当局が車両を埋めたり, 原因究明よりも運行再開を優先しているのは, 何ら不思議なことではなく, 至極
タイバニ，動:伝える，形:null, 例:タイバニの UST 配信と夜勤がかぶらないことをいつも祈っているシスマですこんばんわ…今回は休みの日に配信なので
ジャローダ，動:使える，形:null, 例:ジャローダで両壁貼っても、積み技持ちがズキンしかいなかったことだし抜いた
HTC ハイロード，動:行く，形:null, 例:1 位　マーク・カヴェンディッシュ（イギリス、HTC・ハイロード）　2h27’02”
ピングドラム，動: 見る，形:null, 例:今回のピングドラムは何か起こるらしいですよ〜

れる．

4.2 形容詞

形容詞は取得結果が null となっている．形容詞を取得できなかった要因としては，未知語・同類語を含むの文の中に形容詞が含まれていなかった，形態素解析の段階で形容詞を名詞や固有名詞などと誤って判断をした，あるいは単純に取得するコーパスの量が少なかったことなどが考えられる．形容詞は文章に存在する品詞としては割合が少なく，名詞や動詞と比べて取得できる確率が低い．今後は抽出するテキストの量を増やすなど，実験回数を増やして原因を調査する必要がある．

4.3 例文

例文は未知語を含んだ文をブログ記事の中からランダムで選び，未知語の前後 100 文字程度を出力している．ランダム表示にすることによってバリエーションが増え，未知語理解の助けになると考えられる．

4.4 カテゴリ分け

未知語と単語のカテゴリ分類をする段階で，カテゴリ名に Yahoo ブログカテゴリ，分類方法にシンプソン係数を用いている．動詞・形容詞は未知語と同じカテゴリに分類された同類語から取得するので，同類語を正しく得られなければ最終的な結果にも影響が出る．同類語の推定は今後の課題である．

5 おわりに

本稿では未知語の意味を推定する手法を提案した．

今後の予定として，提案手法の精査と改善，人手による判断を用いた評価実験などが挙げられる．

参考文献

[1] 服部 峻，亀田 弘之，Web テキストにおける未知語の頻度調査，電子情報通信学会技術研究報告. TL, 思考と言語 110(63), 7-12, 2010-05-21

[2] nakatani @ cybozu labs http://labs.cybozu.co.jp/blog/nakatani/2007/09/web_1.html