Vol.2010-SLDM-144 No.50
Vol.2010-EMB-16 No.50
Vol.2010-MBL-53 No.50
Vol.2010-UBI-25 No.50
2010/3/27

Twitterにおけるコンテキストと単語の相関関係分析

荒川 豊[†1] 田頭茂明[†1] 福田 晃[†1]

本研究では，コンテキストアウェア IME 実現へ向けて，コンテキストと入力文字列との相関関係を明らかにするために，Twitter（ツイッター）のつぶやきを収集し分析を行った．ツイッターを分析対象とした理由は，位置情報が付加された文字列が大量に得られることと幅広いユーザ層の文字列が得られることからである．2009 年 12 月 15 日から 2010 年 2 月 1 日の位置情報付きの 13590 件のツイートに対して，位置情報から得られるランドマーク情報と，時間情報から得られるテレビ番組情報とのマッチングを行ない，取得したツイートのうち，4.83%が発言した位置を元に得られるランドマーク情報を含み，8.16%が発言した時間を元に得られるテレビ番組情報を含んでいることを明らかにした．また，一致した文字列は，2〜3 文字であることや Web 検索結果の上位 10 件に約 45%が含まれていることを明らかにした．

Relational Analysis between User Context and Input Word on Twitter

Yutaka Arakawa ,[†1] Shigeaki Tagashira [†1]
and Akira Fukuda[†1]

The objective of this paper is to clear out the relation ship between user's context and really used words in order to realize the context-aware IME. In this paper, we target public tweets of Twitter, because it includes various user's real sentences with geocode (latitude and longitude). We analyze 13590 tweets that have collected from 15 December 2009 to 1 February 2010 for specifying the relationship to landmark information and TV program. As a result, we show that 4.83% of tweets include landmark words, and 8.16% of tweets include TV program words. Additionally, we bring out that average length of concerted words is about 2.5 words, and 45% of them are included in top 10 of web search results.

1. はじめに

近年の調査では，携帯端末からのインターネットアクセスが全体の 5 割を超えており，その 8 割以上が情報を探す際にキャリアが用意したメニューからの選択ではなく，Google 等の検索エンジンに文字列を入力して，情報にアクセスしていることが判明している[1)]．また，ユーザインターフェースの向上と製品のライフサイクルの観点から、iPhone や Google phone など最小限のハードウェアキーしか持たないタッチパネル端末が増加しており，改めて，携帯端末における省入力化への要求が高まっている．これまで，携帯端末における文字入力は，キー入力方式の改善，辞書の拡充，予測変換，学習といったさまざまな手法で省入力化が図られてきている．その中で，近年，iWnn[2)] という携帯向け IME において，電話帳の登録情報や季節・時間帯など，ユーザの利用状況（コンテキスト，と定義）に応じて，予測変換候補を動的に変化させる手法が用いられ注目されている．しかしながら，利用しているコンテキストは端末上で取得可能な情報（ローカルコンテキスト，と定義）に限られており，モバイルコンピューティング環境において特徴的なコンテキストである位置情報が考慮されていないなど，改善の余地も多い．また，これらの手法を用いた場合も，初めて入力する地名やニッチなランドマーク名（例えば，ビル名や交差点の名前，レストラン名）など，辞書データに登録されていない文字列に関しては変換候補として提示できないといった問題もある．

こうした背景から，我々はユーザのリアルなコンテキスト（グローバルコンテキスト，定義）を加味したコンテキストアウェア IME システムを提案している[3)]．グローバルコンテキストとは，ユーザの現在位置，スケジュール情報やプレゼンス情報，現在のニュースや最近の話題といった周りの状況などから推測されるユーザの状態のことである．近年では，多くの携帯端末に GPS が標準搭載されていることと，ニュースなど多くの Web API（Web Application Program Interface）が公開されていること，スケジュール情報やプレゼンス情報なども今後 NGN（Next Generation Network）ではオープン化していくと思われることから，今後ますますこのようなコンテキストサービスが発展していくと考えられる．

最初の段階として，グローバルコンテキストとしてユーザの位置情報を用い，位置を元に得られるランドマーク情報から動的に辞書を生成するコンテキストアウェア IME システム

†1 九州大学大学院システム情報科学研究院
Graduate School of Information Science and Electrical Engineering, Kyushu University

Vol.2010-SLDM-144 No.50
Vol.2010-EMB-16 No.50
Vol.2010-MBL-53 No.50
Vol.2010-UBI-25 No.50
2010/3/27

を構築している．ランドマークとは，地図上の目印となるものであり，駅や役所，学校，病院，郵便局，交差点などの名称のことを指す．我々のシステムを用いることにより，駅の近くでは駅名が優先されたり，同じ「し」でも現在位置により新宿，品川，新橋の順序が変わるといった入力支援機能が追加されるとともに,「九大伊都キャンパス」や「アクトシティ浜松」といった通常の辞書には登録されていない単語を変換候補として表示することが可能となる．しかしながら，このような単語が変換候補として表示されることが，どの程度省入力化に寄与できるのかは定かではなく，駅で乗り換え案内を使うときに駅名が出たら便利に違いないという仮説を元に，これまで研究を進めてきた．文字入力の改善具合を定量的に評価するためには，従来，サンプル文章を入力するのに必要な平均打鍵回数や入力時間などを指標にするのが一般的であったが，コンテキストアウェアなシステムでは，ユーザのコンテキストは多様性が極めて高く，一概に打鍵回数や入力時間で評価することは難しい．そこで，我々は早期にプロトタイプを作成し，実証実験を通じて，省入力化の効果を測定することを試みている．その結果，ある程度の有効性は示すことができたものの，実証実験の被験者数が少ないという問題や，理系学生に偏向しているという問題は払拭されていない．

そこで，さまざまな層のユーザにおいて，より大規模に，提案システムの有効性を検証する方法として，インターネット上で得ることができる膨大な文字列に着目した．我々のシステムではコンテキストとして，位置情報を用いるため，位置情報が付与されている文字列，である必要がある．インターネット上では膨大な文字列を得ることができるが，通常の記事やブログなどには位置情報が埋め込まれることはない．しかしながら，偶然にも昨年 11 月，Twitter 社からつぶやき（ツイート）に対して，位置情報を付与できる Geotagging API が発表され，ツイートに位置情報を付与したり，付与された位置情報を取得することができるようになった．Twitter（ツイッター）とは，140 文字以内のツイートを投稿しあうコミュニケーションサービスであり，近年爆発的に普及している．Twitter のつぶやきは，1 日当たり 5000 万件，2010 年 1 月には月間 12 億件と膨大であり，多種多様なユーザが含まれている．さらに，これらは公開されている API を介して自由に取得することが可能であることから，提案システムの評価に適していると考えた．

本研究では，ツイッターにおける位置情報付きのツイートを 2ヶ月にわたって収集し，時間情報と位置情報に関して，それぞれランドマーク情報，およびテレビ番組表との相関分析を行った．テレビ番組表を用いたのは，iPhone の利用時間の 5 割が自宅からのアクセスという情報に基づき，家でテレビを見ながらツイッターをしている状況が多いのではないかと推測したからである．ランドマーク情報は，Yahoo!ローカルサーチ API からツイートに付与された座標を中心に半径 1 キロ以内の主要なランドマークを取得している．また，テレビ番組表は，東芝が提供している「ネット de ナビ番組表」から番組情報を取得し，Yahoo!日本語形態素解析 API と Yahoo!キーフレーズ抽出 API を用いて，主要な単語を抽出している．これらに関して，まずユーザのさまざまなコンテキストに対してシステムが提供する入力候補に関して分析を行い，次にその入力候補と実際に入力された文字列との関連を明らかにする．

以降では，第 2 章においてこれまで我々が提案しているコンテキストアウェア IME について説明し，第 3 章では Twitter について説明する．第 4 章で，今回行った分析の概要および手法を説明し，第 5 章において分析結果を示す．最後に，第 6 章で本研究および今後の課題を総括する．

2. コンテキストアウェア IME とは

これまでの代表的な省入力化手法としては，1）キー入力方式の改善，2）辞書の拡充，3）予測変換，4）学習，などがあげられる．例えば，1）の例としては，一般的な入力方式であるマルチタップ方式（押す回数で「あ→い→う」と変化する）に対して，子音と母音のツータッチで入力するポケベル方式，入力したい文字が割り当てられているキーを 1 回だけ押し文字列を推測するシングルタップ方式 T9[4]（”1681”と押すと”おはよう”を推薦），入力したい文字が割り当てられているキーを押し，その状態から指を四方にスライドさせることによって入力するフリック方式などがある．2）の例としては，ユーザによる特定単語の登録機能や，外部辞書の追加機能が上げられる．外部辞書の追加機能とは，インターネット用語辞典や人名辞典など，特定の分野に特化した辞書を目的に応じて追加できる機能である．さらに近年では，ネットワークで辞書を共有し，ユーザ全員が単語の登録・共有を行うことのできる Social IME[5] が提案されている．3）の例としては，前方一致検索による全体文字列の推測や，文脈に基づいた助詞・助動詞などの推測があり，POBox[6] を筆頭に，現在広く普及している．4）の例としては，過去のユーザの入力単語を記憶しておき，仮名漢字変換や予測変換での単語候補において，使用頻度と使用履歴に基づいたソートが行われるのが一般的である．特に学習を用いた予測変換は，個人の嗜好を反映しているため，メールの作成などの日常的な文字入力シーンに対して有効なアプローチとなっている．

しかしながら文字入力のシーンは多様化してきており，メールやメモの作成のみならず，乗り換え案内の利用や周辺情報の検索なども携帯端末上で行うようになってきた．このような多様な文字入力シーンに対しては，使用頻度や使用履歴という指標の一律な適用だけで

Vol.2010-SLDM-144 No.50
Vol.2010-EMB-16 No.50
Vol.2010-MBL-53 No.50
Vol.2010-UBI-25 No.50
2010/3/27

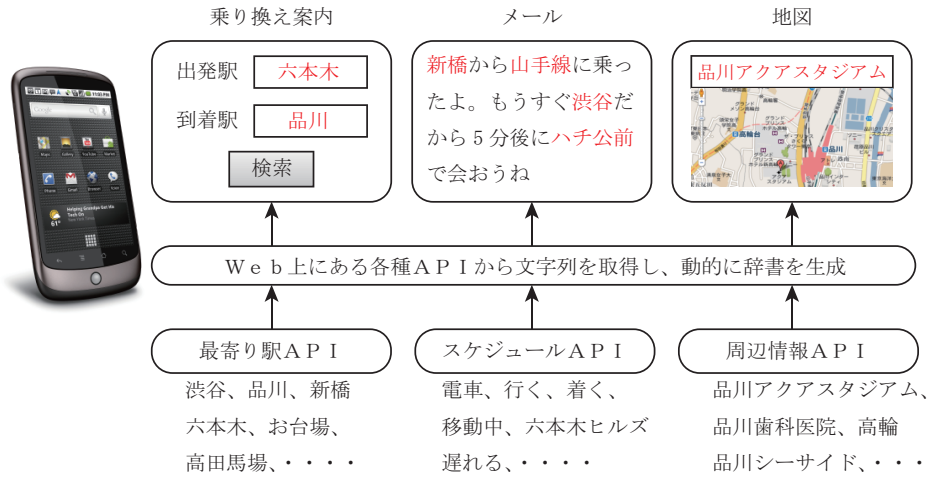

図 1 コンテキスト IME が実現したいサービスの例

は十分に対応することができない．そこで近年では，ユーザの状態（コンテキスト）を推定し，より入力時の状態に即した単語を推薦する研究が行われている．携帯端末向け IME である iWnn[2] では，電話帳の登録情報や季節・時間帯などを利用し，予測変換候補を動的に変化させることで省入力化を支援する仕組みが実装されている．しかしながら，利用しているコンテキストは端末上で取得可能な情報（ローカルコンテキストと定義）に限られており，モバイルコンピューティング環境において特徴的なコンテキストである位置情報が考慮されていないなど，改善の余地も多い．また，これらの手法を用いた場合も，初めて入力する地名やニッチなランドマーク名（例えば，ビル名や交差点の名前，レストラン名）など，辞書データに登録されていない文字列に関しては変換候補として提示できないといった問題もある．

そこで我々は，携帯端末における新たな省入力化へのアプローチとして，位置情報やネットワークを介して得られるプレゼンス情報（グローバルコンテキストと定義），さらにそれらから副次的に得られる周辺情報（ランドマーク名，最寄り駅名，レストラン名）などを考慮し，ユーザのいる場所・時間・状態に応じて，より適した文字列を推薦するコンテキストアウェア IME システムを提案している[3]．図 1 にコンテキスト IME が実現したいサービスの例を示す．例えば，乗換検索を行うと，近くの駅名が予測変換候補として出てきたり，

図 2 Android の OpenWnn 上への実装画面（位置は浜松駅）

旅行にいくと，その場所付近の観光名所が既に辞書に入っていたり，「し」で始まる「新宿」「渋谷」「新橋」などが位置によりソートされていたりといった効果を狙っている．それを実現するための手法として，携帯端末に搭載された GPS センサや加速度センサ，地磁気センサなどから，ユーザの位置や移動方向を取得するとともに，ネットワークを通じて周辺情報やスケジュール情報，プレゼンス情報などを取得し，ユーザのコンテキストを推定する．次に，コンテキストに基づいて，ネットワーク上のさまざまな Web API から単語を取得し，動的にコンテキスト辞書を更新する．通常，ある辞書を作る際には初期コストやメンテナンスコストなどの膨大な人的コストが必要とされるが，Web API から提供されるデータを活用することで，コスト負担なしに辞書作成を行うことができる．さらに，全ての Web API を辞書全体と見なせば，辞書内の単語は Web API の種類によりクラスタリングされており，Web API へのクエリによってフィルタリング可能であるといえるため，詳細かつ的確な単語が取得可能であると考えられる．このようにして作成された辞書内の単語をもとに，予測変換候補としてユーザに提示することで，状況に応じた単語の推薦を実現する．さらに，Web API から取得した単語のソート，及び推定と決定の繰り返しによる学習フィードバックを取り入れ，個々のコンテキストと文字列を関連づけていくことにより，パーソナライズされた日本語入力システムを実現する．

我々は，提案の有効性を検証するために，市販の IME である ATOK 上にダイレクトプラグインとして実装したプロトタイプ，および Android 上の OpenWnn を拡張したプロトタイプ 2 を作成した[7]．特に後者は，提案システム以外にも，GPS のロギングソフトウェア，IME の変換履歴保存スクリプトを実装し，九州大学の学生に日常生活で利用してもらっ

Vol.2010-SLDM-144 No.50
Vol.2010-EMB-16 No.50
Vol.2010-MBL-53 No.50
Vol.2010-UBI-25 No.50
2010/3/27

た．しかしながら，メール内容にはプライバシー情報が含まれること，普段利用しているメールアドレスが利用できないことなどの理由から，サンプルとして得られた文字列情報は非常に少なく，有効性を検証するには不十分であった．

3. Twitter に関して

Twitter（ツイッター）[8] とは，2006 年 7 月に Obvious 社が開始した，ユーザが 140 文字以内で「つぶやき（ツイート）」を投稿することで，メールやメッセンジャーよりも，ゆるいつながりを発生させるコミュニケーションサービスである．ツイートは，基本的には誰からも閲覧できる状態（隠すことも可能）である．また，閲覧を申告することをフォローとよび，フォローしているユーザのツイートは，タイムライン，呼ばれるツイート一覧に，ほぼリアルタイムに表示される．

ツイッターは，各種 API がユーザに対して公開されており，サードベンダーを含め，一般ユーザがツイッターと連携したアプリケーションを作りやすい環境が用意されている．さらに，iPhone など，ツイッターに適したスマートフォンの普及が追い風となり，近年爆発的にユーザが増加している．

ツイッターの開発は現在も続いており，昨年の 11 月には Geotagging API が公開された．Geotagging API とは，ツイートに対して，つぶやいた場所の位置情報（緯度・経度）を付与できる API である．これまでも大まかな位置をユーザのプロファイルとして登録することはできたが，つぶやくときに更新されるわけではなかった．Geotagging API のリリースにより，GPS を搭載した iPhone などの携帯端末でつぶやくことによって，付与される位置情報の精度が飛躍的に向上しており，地図と連携したサービスなど新たなサービス領域が生まれつつある．当初は，対応アプリケーションが少なかったため，位置情報が付与されたツイートは少なかったが，徐々に有名なクライアントソフトウェアが対応を進め，現在ではそれなりの数を収集できるようになっていおり，さまざまな位置，さまざまな時間における，さまざまなユーザの入力文字列を容易に入手することが可能となった．そこで我々は，位置情報が付与されたツイートを分析することにより，これまでの研究で前提としてきた，コンテキストと入力単語には相関があるはずという前提の妥当性を検証することができるのではないかと考え，本研究に着手した．

4. ツイート分析の概要

図 3 に示すように，1 つのツイートから，つぶやいた時刻，つぶやいた位置，つぶやいた

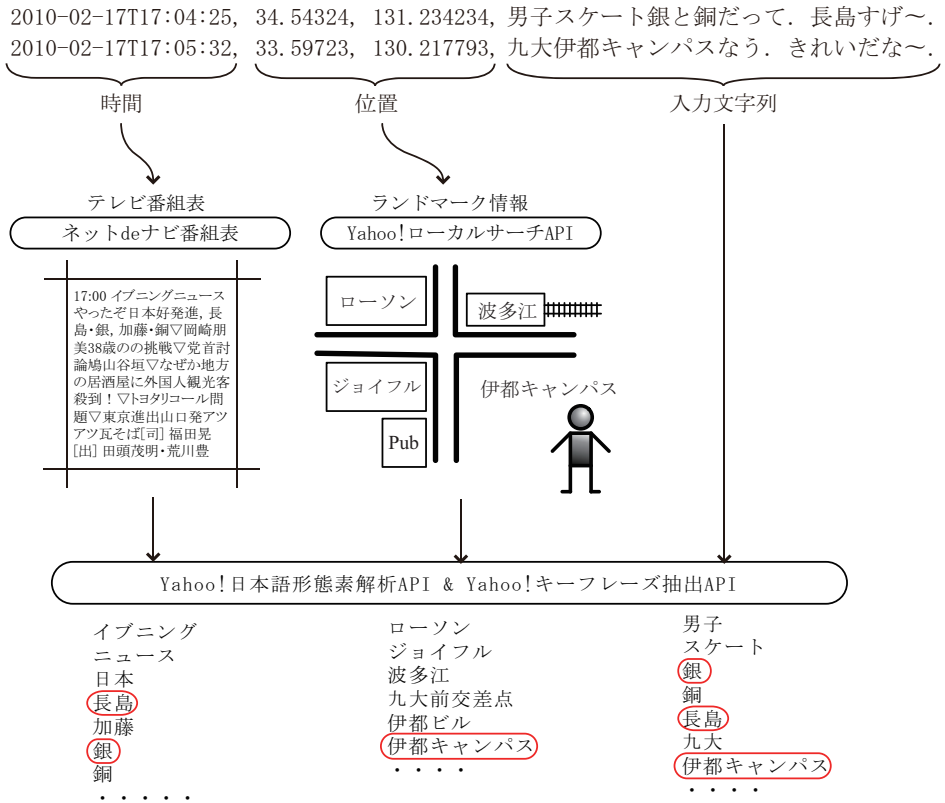

図 3 ツイートから得られる情報とその分析の流れ

文字列，の 3 つの情報を得ることができる．本研究では，ユーザのコンテキストとして，文字を入力した時刻と文字を入力した場所を想定し，そのコンテキストから得られる文字列と，実際に入力された文字列との相関を検証する．コンテキストから得られる情報として，さまざまなものが考えられるが，今回は，時刻情報から得られる情報としてはテレビ番組表内の文字列，位置情報から得られる情報としては周辺のランドマーク名を対象とした．以下に，位置情報が付与されたツイートの取得，テレビ番組表の取得，ランドマーク情報の取得，さらにそれらの相関分析に関して示す．

Vol.2010-SLDM-144 No.50
Vol.2010-EMB-16 No.50
Vol.2010-MBL-53 No.50
Vol.2010-UBI-25 No.50
2010/3/27

4.1 位置情報が付与されたツイートの取得

Twitter API にはさまざまな API が公開されており，キーワード検索やユーザ ID 指定検索などを行うことができるが，位置情報が付与されたツイートだけを取得する API は存在しない．そこで，取得可能な全てのツイートを収集し，日本語かつ位置情報が付与されている発言だけをデータベースに記録するというアプローチを用いている．全ツイートの取得は，昨年 4 月からアルファテストが開始され，この 1 月に正式リリースされた Streaming API を用いる．Streaming API は，Public Timeline と呼ばれる鍵のかかっていない全てのツイートを取得可能な”firehose”, 全てのツイートからランダムにサンプリングされたツイートを取得可能な”gardenhose”，gardenhose の数分の 1 のツイートを取得可能な”spritzer”の 3 種類のレベルがある．spritzer レベルは誰でも利用できるが，firehose レベルと gardenhose レベルは Twitter 社に申請して利用許可を得る必要がある．今回はより多くのツイートを取得するために，Twitter 社に申請し，gardenhose レベルを利用した．ちなみに，firehose レベルは，一般的に利用許可を得るのは難しいとされている．

今回利用した gardenhose レベルでは，全ツイートの約 1/5 程度が得られるとされているが，その中で日本語かつ位置情報が付与されたツイートに絞り込むと予想以上に得られるツイートは少なかった．これは，日本で普及している携帯電話およびクライアントソフトウェアが位置情報の付与に向いていないという問題が考えられる．iPhone などのスマートフォンでは，Geotagging API に対応したクライアントソフトウェアが多数存在するが，携帯電話や PC のブラウザ経由で Twitter にアクセスする場合は，Geotagging API を用いて位置情報を付与することができない．正確には，携帯電話でも，ツイート内に座標を文字列（140 文字のツイートの一部）として書くことは可能であるが，GPS 情報を取得して貼り付けるという作業が必要なことから普及には至っていない．

Geotagging API を用いてツイートに位置情報を付与するためには，クライアントソフトウェアが Geotagging API に対応している必要がある．言い換えれば，位置情報が付与されたツイートを発言した人は，位置情報付与に対応したクライアントソフトウェアを利用しており，その後も同じクライアントソフトウェアを使う限り，位置情報付きのツイートを発言している可能性が高い．そこで我々は，Streaming API から得た位置情報付きのツイートのユーザ ID を用いて，別途 Twitter Search API にアクセスし，同一ユーザの過去の発言をさかのぼって取得するように拡張した．その結果，日本語の位置情報付きツイートを比較的大量に取得することが可能となった．ただし，Twitter Search API では，過去にさかのぼって取得できる発言数に制限があるため，拡張以前のツイートに関してはほとんど収集量を増やすことはできなかった．

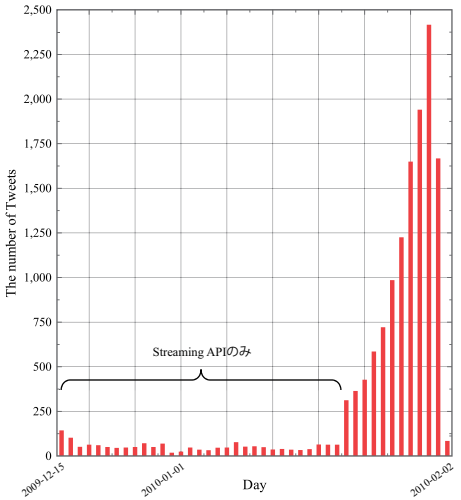

図 4 今回収集したツイート数の日ごとの分布

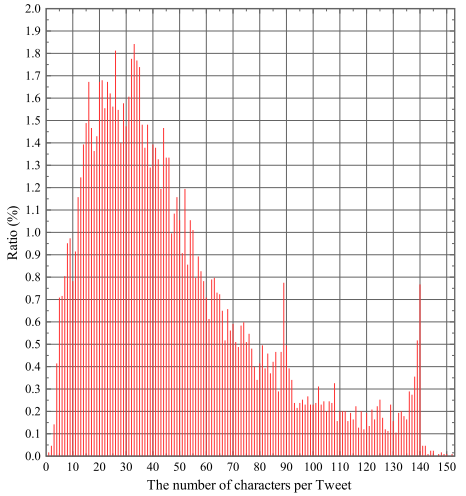

図 5 今回収集したツイートの文字列長の分布

4.2 テレビ番組情報の取得

ソフトバンクの孫社長によると，iPhone 全体のアクセスの内、実に 5 割が自宅の無線 LAN を経由しているそうである．そこで，自宅でつぶやく場合，テレビ番組に関連した文字列，例えば番組名や出演者名を入力することが多いのではないかと考えた．テレビ番組情報は，東芝が提供している「ネット de ナビ番組表」[9] から取得している．放送される番組の概要や出演者名を含んでおり，それらを「Yahoo!日本語形態素解析 API」と「Yahoo! キーワード抽出 API」という 2 種類の API を用いて，単語に切り分ける．より正確に検証するためには，発言された位置で放映されているテレビ局を選択する必要があるが，緯度・経度から放送されているテレビ局を探すことは困難であることから，今回は在京局（日本テレビ，フジテレビ，NHK 総合，NHK 教育，TBS テレビ，テレビ東京，TOKYO MX）と福岡の地方局（RKB，九州朝日放送，テレビ西日本，福岡放送，TVQ 九州）の合計 12 局を収集対象とした．テレビ番組データは過去にさかのぼって収集することが困難であるため，今回は収集スクリプトが完成した 1 月 7 日以降のデータが対象となる．

4.3 ランドマーク情報の取得

ランドマーク情報とは，駅，役所，学校，病院，郵便局など地図上で目印となる情報であ

Vol.2010-SLDM-144 No.50
Vol.2010-EMB-16 No.50
Vol.2010-MBL-53 No.50
Vol.2010-UBI-25 No.50
2010/3/27

る．本研究では，「Yahoo!ローカルサーチ API」を用い，ツイートの座標から 1 キロ以内に存在するランドマーク情報を最大 100 件取得する．そして，それらをテレビ番組情報と同様に，Yahoo!日本語形態素解析 API と Yahoo!キーワード抽出 API という 2 種類の API を用いて，単語に切り分ける．

5. 結　　果

本論文で取り扱うデータは，2009 年 12 月 15 日 21 時 16 分 09 秒から 2010 年 2 月 2 日 9 時 10 分 5 秒に得られたものである．

5.1 収集したデータの分析

まず，収集できた各種データ（ツイート，ランドマーク情報，テレビ番組情報）に関して報告する．この期間に得られた位置情報が付与された日本語のツイートは 13590 件であり，日ごとの分布は図 4 のようになっていた，当初は，Streaming API による収集だけであり，1 日当たり 50 件〜100 件程度のツイートが得られている，その後，1 月に入り，飛躍的に収集ツイート数が増大していることがわかる．これは，位置情報付きツイートを発しているユーザに対して，Search API を用いて過去の発言を収集する手法を導入したためである．なお，現在もスクリプトは稼働中であり，すでに 40000 件超のツイートが収集されている．

今回，分析対象となる 13590 件のツイートの平均文字列長は，48.22 文字で，その分布は図 5 のようになっている．ツイッターの 140 文字という制限に対して，30 文字程度のツイートが多いことがわかる，つまりメールやブログなどと異なり，隙間時間で投稿する “つぶやき” は比較的短い文章が多いという結果である．また，140 文字を超えるツイートも数件見受けられるが，これは，「“（クォーテーション）」や「&（アンバサンド）」が HTML エンコーディング処理により，「"」や「&」に置換されているためである．

ある座標において，Yahoo!ローカルサーチ API から得られるランドマーク数の平均は 28.1 件であり，分布は図 6 のようになっている，20 件という結果が突出しており，最大で 66 件得られた座標もあった．一方，座標によっては 0 件という場合も 5.5%程度見られた．

同様に，ある時間帯（1 時間）において，テレビ番組表から得られる文字列数の平均は 207.2 個であり，分布は図 7 のようになっておる．図 7 は，ちょうど 50 個，100 個というわけではなく，50 単位でサンプリングした結果であり，50 は 0 個〜49 個，100 は 50 個〜99 個の表している．図より，150 個前後が多いが，中には 800（750 個〜799 個）という時間帯もあり，ランドマーク情報と比較して，得られる情報が約 10 倍近くあることがわかった．

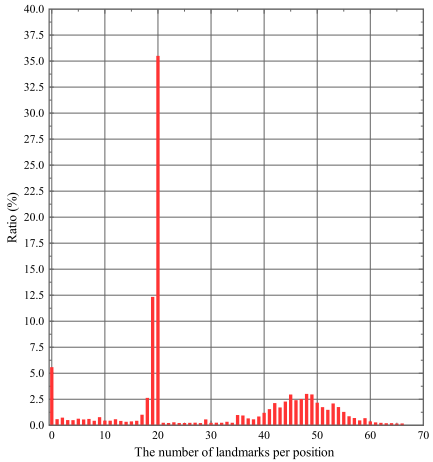

図 6 Yahoo!ローカルサーチ API から得られるランドマーク数の分布

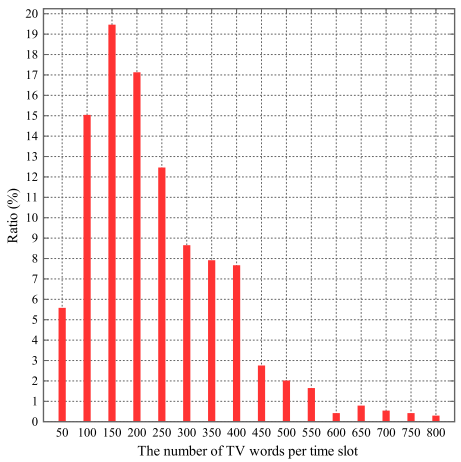

図 7 テレビ番組表から得られる 1 時間当たり単語数の分布

5.2 マッチング分析

次に，収集したデータを用いて，実際に入力された文字列と，位置や時間を元に得た文字列との相関関係を分析する．まず，ツイートを発した位置に元に Yahoo!ローカルサーチ API から得られるランドマーク情報を，ツイート自身に含んでいた割合（含有率）は，4.83%（13590 件中 656 件）であった．一方，テレビ番組表から得られた文字列を，ツイート自身に含んでいた割合は，8.16%（13590 件中 1109 件）であった．この数字は，少ないようではあるが，20 ツイートに 1 ツイートは，周辺のランドマーク情報を含んでいることは事実であり，検索候補の絞り込みアルゴリズム次第では有用であると考えられる．

具体的に含まれていた文字列の上位 10 件を表 1 に示す．これを見ると，きわめて有名な地名が多いことがわかる．さらに，本提案では，「伊都キャンパス」や「キャナルシティ博多」といったある程度長い文字列が出現することを想定していたが，予想外に短い単語が利用されることが多いこともわかった．一致した文字列長の平均は，ランドマーク情報から得られた文字列が平均 2.56 文字，テレビ番組表から得られた文字列が平均 2.3 文字であり，それぞれの分布は図 8 となっており，ほとんどの文字列が 4 文字以下であることがわかる．この結果は，4 文字以下の文字列だけを推薦すればいいとも取れるが，現在の携帯端末における文字入力では長い文字列を入力しにくいため，結果として短い単語ばかり利用されてい

Vol.2010-SLDM-144 No.50
Vol.2010-EMB-16 No.50
Vol.2010-MBL-53 No.50
Vol.2010-UBI-25 No.50
2010/3/27

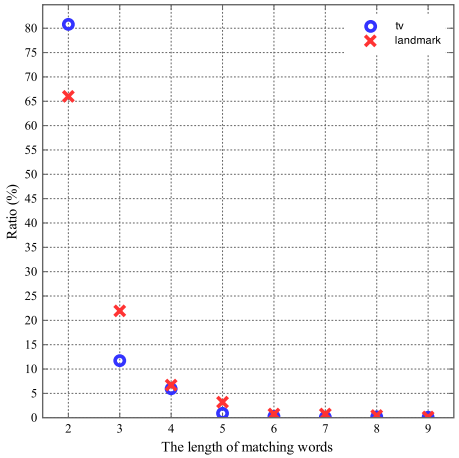

図 8 一致した文字列の長さの分布

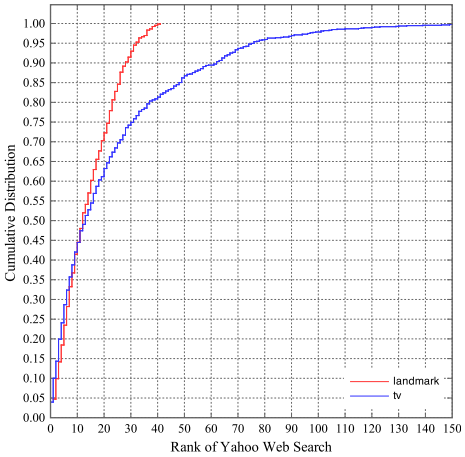

図 9 一致した文字列の Yahoo 検索を用いた順位付けの累積密度分布

ると考えることもできる．

最後に，現在，得られた文字列の絞り込みやソート手法との一つとして，web 検索結果の総数の利用を考えている．そこで，今回の得られた文字列に関して，実際に一致した文字列が，そのときの候補の中で何位に位置していたかを分析した．今回は，Yahoo!検索 API を用い，入力候補すべてに関して Yahoo!検索のヒット数を取得した．その累積密度分布を図 9 に示す．この図から，10 位以内に含まれる確率が約 45%であり，得られた文字列の中から，web 検索結果の総数に基づき絞り込む手法は有用性が高いと考えられる．

6. おわりに

本研究では，コンテキストと単語の相関関係を明らかにすることを目的として，ツイッターを対象に，ツイートの位置情報や時間情報から得られる文字列の分析，および実際に入力との相関関係を分析した．2009 年 12 月 15 日以降，日本語かつ位置情報が付与された 13590 件のツイートを収集した結果，取得したツイートのうち，4.83%が発言した位置を元に得られるランドマーク情報を含み，8.16%が発言した時間を元に得られるテレビ番組情報を含んでいることを明らかにした．また，一致した文字列は，2~3 文字であることや Web 検索結果の上位 10 件に約 45%が含まれていることを明らかにした．

表 1 一致した文字列の上位 10 件

	ランドマーク	テレビ番組表
1	東京	今日
2	渋谷	日本
3	新宿	いま
4	横浜	もの
5	川崎	情報
6	ビル	世界
7	立川	東京
8	大阪	ニュース
9	日本	明日
10	名古屋	こと

謝辞 本処理系の開発, 及び検証は, 日本電信電話株式会社 NTT サービスインテグレーション基盤研究所と国立情報学研究所の提供する研究設備, 回線を利用した共同研究の一環として実施している．ここに記して謝意を示す．

参 考 文 献

1) rTYPE: 「ネットは PC より携帯」携帯ネット歴 5 年以上では半数以上 — rTYPE アイシェアオンラインリサーチサービス 市場調査公開 (2009). http://release.center.jp/2008/11/0502.html.
2) オムロンソフトウェア株式会社：iWnn. http://www.omronsoft.co.jp/SP/.
3) 末松慎司，荒川 豊，田頭茂明，福田 晃：ネットワークを用いたコンテキストアウェア日本語入力支援システムの提案，信学技報，NS2009-136, Vol.109, No.326, pp.89–94 (2009).
4) Grover, D., King, M. and Kuschler, C.: Patent No.US5818437,Reduced keyboard disambiguating computer, Tegic Communications, Inc., Seattle, WA (1998).
5) 奥野 陽，萩原将文：インターネットを用いた日本語入力システム，情報処理学会第 190 回自然言語処理研究会 (2009).
6) Masui, T.: *POBox: An Efficient Text Input Method for Handheld and Ubiquitous Computers*, Lecture Notes in Computer Science, Vol.1707/1999, pp.289–300, Springer Berlin / Heidelberg (1999).
7) 荒川 豊，末松慎司，田頭茂明，山口雄輔，田中裕大，福田 晃：[技術展示] ネットワーク連携コンテキストアウェア日本語入力支援システムの実装，信学技報，MoMuC2009-58, Vol.109, No.380, pp.31–34 (2010).
8) Twitter 社：Twitter. http://twitter.com/.
9) 東芝：ネット de ナビ番組表. http://tvsurf.jp/tv/.