3F3-4in

# Web 上のライフストリームからのユーザ行動情報の抽出

## Extraction of human activity information from Web lifestream

佐々木 健太[*1] 長野 伸一[*1] 長 健太[*1] 川村 隆浩[*1]
Kenta Sasaki Shinichi Nagano Kenta Cho Takahiro Kawamura

[*1] 株式会社東芝研究開発センター
Corporate Research & Development Center, Toshiba Corporation

We propose a method of extracting human activity and attribute information from blog articles focused on outdoor activity, which are main resources of Web lifestream, utilizing a human activity verb dictionary. We define human activities as only human-centric actions, such as eat, play, think and so on, and define attributes as information associated with human activities. We construct the verb dictionary based on the verb argument structure thesaurus. As a result over 75 % verbs, extracted among 100,000 blog articles, were contained within the verb dictionary and both human activity and attribute information could be correctly extracted to some extent. One of the tasks for the future is to systematize the human activity and attribute information like ontology or semantic network, and summarizing a series of activities or estimating the past activities would be possible.

## 1. はじめに

近年，ユーザの行動を記録，蓄積するライフログが注目を浴びている．ライフログは様々なサービスに利用可能であり，集合知として利用すれば人間の行動予測といった近未来的サービスの実現も夢ではない [Sellen 07]．一方，ライフログを手動記録するのは非常に手間である．そこで，自動取得するための情報源として，ライフストリームと呼ばれる Web データ（ブログ，Twitter）が挙げられる．ライフストリームとは，そのユーザに関する情報が絶え間なく流れる様子を指し，テキスト形式で表現されることが多い．ところが，これらにはユーザの様々な行動が含まれ，一般的事実やユーザとは直接関係のない出来事なども多く含まれる [Kwak 10]．ライフストリームからライフログを自動取得するには，ユーザの行動だけを抽出した上で，それらの情報を整合することが重要となる．これより，本研究では人間が取りうる行動の体系化を目的とし，それに向けた取り組みの一環として，ライフストリーム，特にブログから行動，行動属性を抽出する手法を提案する．なお，本研究では，人間が主体となる動作を行動，また行動に付随する情報を行動属性と定義する．

2 章で問題定義，3 章で提案手法，4 章で評価結果，5 章で関連研究を述べる．最後に，6 章でまとめを述べる．

## 2. 問題定義

本研究では，ブログ，Twitter のようなテキスト情報からユーザの行動，行動属性を抽出する問題について述べる．テキスト情報からユーザの行動，行動属性を抽出するためには，何かしらの手掛かりが必要である．そして，主に動作や状態を表し，主語，目的語などの項を伴うことが多い動詞が，その重要な手掛かりの一つとなる．ところが，動詞は「歩く」「悩む」のように人間が主体となる行動動詞と，「発車する」「積もる」のように人間が主体とならない非行動動詞に分類することができる．さらに，前者は，「歩く」のように客観的に観察可能な外面的行動動詞と，「悩む」のように心理的な行動を表す内面的行動動詞に分類することができる（図 1）．ブログから行動，行動属性を自動抽出するためには，行動動詞を手掛かりとして持ち合わせておく必要がある．

連絡先：佐々木健太，kenta.sasaki@toshiba.co.jp

ところが，行動動詞には外面的行動動詞，内面的行動動詞など様々なものが存在し，これらをゼロから漏れなく列挙することは非常に困難である．そこで，本研究では，自然言語処理用の辞書に基づいて行動動詞判定辞書を構築し，構築した辞書を利用することでブログから行動，行動属性を抽出する．

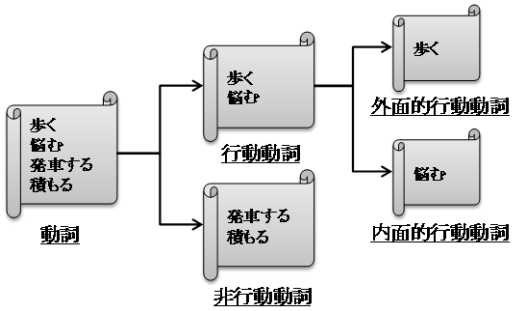


図 1 動詞分類

## 3. 提案手法

本章では，提案手法を紹介する．提案手法，特に行動抽出の流れを図 2に示す．最初に，自然言語処理用の辞書から行動動詞判定辞書を事前に構築しておく．そして，外出状況を綴ったブログを Web から収集した後に，本文だけを抽出し，形態素解析にかける（図 2(a),(b),(c)）．最後に，形態素解析結果に対し，行動動詞判定辞書を照らし合わせることで，行動だけを抽出する（図 2(d)）．

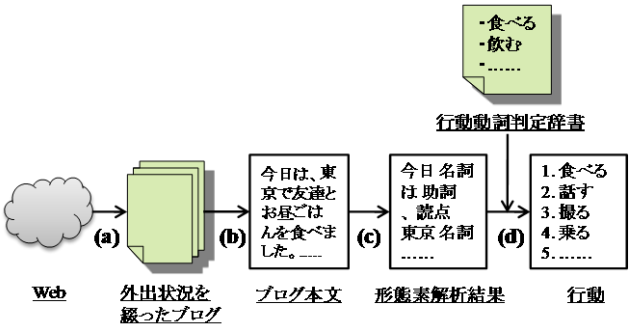


図 2 行動抽出の流れ

以下，行動動詞判定辞書構築，行動抽出，行動属性抽出について詳細に述べる．

### 3.1 行動動詞判定辞書構築

行動動詞判定辞書構築には，岡山大学の竹内講師らが開発した自然言語処理用の辞書である，動詞項構造シソーラスを利用する [竹内 08]．一般に，動詞は複数の意味を持つが，係り関係のある名詞との組み合わせによって意味が特定される．すなわち，各動詞は「本，を，読む」のように深層格（対象主，動作主），表層格（ヲ，ガ），動詞の組み合わせで，動詞の意味が異なる．このような関係をすべての名詞，助詞，動詞の組で記述することは困難であるが，動詞の内部の意味構造を分解して，深層格のレベルで有限個の語義へ対応付け，5 階層に分類している．複数の語義が存在する動詞は，その語義の数だけ辞書に登録されている．例えば，「降りる」という動詞は，乗降の「飛行機，を，降りる」と，自然現象の「霜，が，降りる」などの 9 つの語義が登録されている．2010 年 9 月 15 日時点では，動詞 5,170 語（10,359 語義）が格納されており，合計 1,154 カテゴリに分類されている．一般の自然言語処理用の辞書には，動詞に関しては品詞の分類に関する情報が登録されているにすぎないのに対して，動詞項構造シソーラスでは，動詞は内部の意味構造をもとに分類されている．この意味構造を参照することで，行動動詞か非行動動詞かの判断が容易になると考え，行動動詞判定辞書構築のベースとして利用する．

具体的には，まず，動詞項構造シソーラスに格納されている各語義に対して，その動詞が行動動詞になるのか，ならないのかを分類した．そして，1 つの語義でも行動動詞になると分類した動詞を行動動詞，それ以外の動詞を非行動動詞と判定し，行動動詞判定辞書に記録した．前記の例では，「霜，が，降りる」の「降りる」は行動動詞にならないが，「飛行機，を，降りる」の「降りる」は行動動詞になるので，「降りる」という動詞を行動動詞と判定した．なお，本研究では直接利用しないが，行動動詞をさらに外面的行動動詞と内面的行動動詞に分類した．

### 3.2 行動抽出

外出状況を綴ったブログから，行動動詞判定辞書を利用して行動を抽出する．外出状況を綴ったブログを対象とした理由は，行動動詞が出現する可能性が最も高く，提案手法の有効性を確認するには最適であると考えたからである．

最初に，「浜松町」「川崎」などの地名を検索キーワードとした Google ブログ検索*により，約 10 万件のブログを収集した（図 2 (a)）．地名を検索キーワードとすることで，外出状況を綴ったブログを優先的に収集することができると考えた．次に，収集したブログから本文を抽出した（図 2 (b)）．ここでは，Perl で公開されている，HTML::ExtractContent というモジュールを利用した．これは，タグを手掛かりにしたヒューリスティクスな方法で，HTML から本文を抽出するモジュールである．そして，JUMAN† を利用して，ブログ本文を形態素解析にかけた（図 2 (c)）．JUMAN を利用した理由は，形態素の品詞情報だけでなく，カテゴリ情報，ドメイン情報も出力するため，行動属性抽出，さらに行動情報を体系化する場合に有用であると考えたからである．最後に，形態素解析結果から動詞だけを取り出し，行動動詞判定辞書を参照することで動詞だけを抽出した（図 2 (d)）．これにより，形態素解析結果から行動動詞だけを抽出することができる．なお，「出社，する」「撮影，する」のような，「（サ変名詞），する」も動詞として抽出した．行動動詞判定辞書には，これらの動詞は初めから格納されている．また，「ある」「する」「なる」などの平仮名 2 文字の動詞は出現回数が多くノイズとなるため，ストップワードとした．以上の処理により，外出状況を綴ったブログから，行動だけ（ライフログ）を抽出することができる．

### 3.3 行動属性抽出

行動動詞判定辞書だけでなく，特定の名詞，助詞，動詞の組み合わせを抽出するパターンマッチングを利用して，行動，および行動属性を抽出する．各行動に対して，「何を（what）」，「どこで（where）」，「いつ（when）」，「誰と（whom）」，「どのように（how）」に相当する情報を，行動属性と定義した．

具体的には，ブログ収集，本文抽出，形態素解析までは，行動抽出と同一の処理を行った．そして，形態素解析結果から行動動詞を抽出する際に，行動動詞判定辞書から行動動詞と判定され，かつ属性毎に事前に用意したパターン（表 1）と一致する名詞，助詞，動詞の組み合わせのみを抽出した（図 3）．行動動詞判定辞書から行動動詞と判定されても，一致するパターンが存在しない動詞は抽出しなかった．ここで，抽出した動詞を行動，名詞を行動属性とした．この際，文字列パターン以外に，名詞の品詞細分類，およびカテゴリを抽出条件として利用した．品詞細分類には「（普通名詞）」，「（サ変名詞）」，「（固有名詞）」などの詳細な品詞情報，カテゴリには「（場所）」，「（人）」，「（人工物）」などの意味情報が付与されている．「弟，で，待つ」の様な，文法的に有り得ない名詞と動詞の組み合わせを抽出するのを防ぐため，品詞細分類，およびカテゴリによる条件付けを加えた．以上の処理により，例えば，「食べる」という行動に対して，「what: お昼ごはん」，「where: 東京」，「whom: 友達」という属性を抽出することができ，行動に関する詳細な情報を得ることが可能となる．

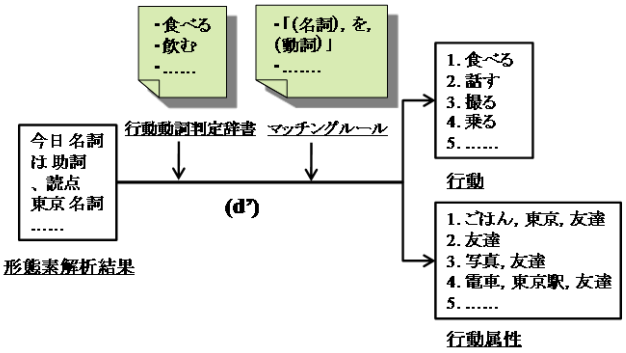


図 3 行動属性抽出の流れ

表 1 行動属性抽出パターンマッチングルール

| 行動属性 | パターン | 名詞条件 |
|---|---|---|
| What | （名詞）を（動詞） | 細分類（普通名詞，固有名詞，地名，人名，数詞） |
| Where | （名詞）で（動詞） | 細分類（地名），またはカテゴリ（場所） |
| When | （名詞）（動詞） | カテゴリ（時間） |
| Whom | （名詞）と（動詞）<br>（名詞）から（動詞）<br>（名詞）に（動詞）<br>（名詞）を（動詞）<br>（名詞）へ（動詞） | 細分類（人名），またはカテゴリ（人） |
| How | （名詞）で（動詞） | カテゴリ（人工物） |

* Google ブログ検索: http://blogsearch.google.co.jp/

† JUMAN: http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

## 4. 評価

本章では，行動動詞判定辞書を利用した行動抽出，および行動属性抽出，それぞれの評価方法，結果について述べる．

### 4.1 評価方法

行動抽出手法の妥当性を評価するため，形態素解析結果からの行動動詞抽出（図 2 (d)）に着目した．ここでは，行動動詞判定辞書を参照することで，各動詞は行動動詞，非行動動詞，未分類動詞の 3 種類に分類される．行動動詞判定辞書に格納されていない，つまり動詞項構造シソーラスに格納されていない動詞を未分類動詞と定義した．未分類動詞は，それが行動を表しているのか，いないのか判別できない．そのため，未分類動詞に分類される動詞が少なければ少ないほど，本提案手法の有効性が高いということが言える．以上のことより，各動詞の分類結果から，行動抽出の評価を行った．同様に，行動属性抽出手法に関しても，抽出した各動詞の分類結果から評価を行った．

### 4.2 行動抽出の評価

行動抽出における動詞分類結果を表 2，また出現頻度上位 10 位以内の動詞の分類結果を表 3に示す．表 2，表 3より，外出状況を綴ったブログに登場する 75%以上の動詞，および出現頻度上位 10 位以内の全ての動詞を行動動詞か非行動動詞に分類できることを確認した．これより，行動動詞判定辞書を利用した行動抽出手法の妥当性を，一定程度確認することができた．一方，未分類動詞と分類されたものは，(1)「いただく」「わかる」などの平仮名のみで構成されるもの，(2)「込む」「換える」などの動詞の断片として抽出されたもの，および(3)「観る」「仕分ける」などの行動動詞として登録されているべきものの，大きく 3 つに分類することができた．(1)は，さらに，形態素解析誤りによるものと，本来漢字で表現されるべきものに分けられる．前者は，ブログのような砕けた日本語を対象にすることによる形態素解析誤りである．既存の形態素解析エンジンの多くは，ニュース記事のようなフォーマルな日本語を想定しており，ブログや Twitter のような砕けた日本語への利用は想定していない．後者は，本来漢字で表現されるべきものであり，その大部分は漢字で表現したものが行動動詞判定辞書に格納されている．平仮名を漢字に変換することで，これらの動詞を抽出することは可能である．(2)は，「作り込む」「買い換える」などの複合動詞が分割され，その断片として出力されたものである．キーワードとして適切な塊になるまで形態素を結合させる複合語処理により，これらの動詞が分割して抽出されるのを防止することができる．(3)は，そのまま行動動詞判定辞書に追加することで，抽出可能な動詞である．

表 2 動詞分類結果（行動抽出）

| 分類 | 出現頻度（延べ数） | 割合（%） |
|---|---|---|
| 行動動詞 | 2,113,568 | 72.9 |
| 非行動動詞 | 166,980 | 5.7 |
| 未分類動詞 | 619,741 | 21.4 |

表 3 出現頻度上位 10 位以内の動詞（行動抽出）

| 動詞 | 分類 | 出現頻度（延べ数） |
|---|---|---|
| 思う | 行動動詞 | 84,313 |
| 見る | 行動動詞 | 58,614 |
| 行く | 行動動詞 | 53,540 |
| 言う | 行動動詞 | 52,106 |
| 出る | 行動動詞 | 28,907 |
| 感じる | 非行動動詞 | 28,259 |
| 考える | 行動動詞 | 23,877 |
| 書く | 行動動詞 | 21,067 |
| 使う | 行動動詞 | 20,261 |
| 入る | 行動動詞 | 20,157 |

### 4.3 行動属性抽出の評価

次に，行動属性抽出における動詞の分類結果を表 4，また属性毎の出現頻度上位 5 位以内の動詞の分類結果を表 5（what），表 6（where），表 7（when），表 8（whom），表 9（how）に属性と共に示す．表 4より，行動動詞判定辞書に加えてパターンマッチングを利用することで，90%以上の動詞を行動動詞か非行動動詞に分類できることを確認した．パターンマッチングを利用せずに動詞だけを抽出する場合と比較すると，10 ポイント以上向上している．これは，文字列パターン，および名詞の品詞細分類やカテゴリを考慮することで，形態素解析誤りによる動詞の誤抽出が減ったためであると考えられる．また，表 5，表 6，表 7，表 8，表 9より，「写真，を，撮る」「コンビニ，で，買う」「地図，を，見る」のように，外出状況で頻出する行動とその属性を一定数抽出できることが分かった．一方，「物件，を，探す」「町，で，探す」のように外出状況で普段あまり出てこない行動とその属性を抽出してしまっている，また「町，で，探す」「日，行く」のように属性が本来の属性情報の断片として抽出していると思われる場合があることが分かった．前者は，地名を検索キーワードとしてブログを収集したために，不動産などの広告ブログが多数混じってしまい，結果的に普段あまり出てこない行動を抽出したことが原因であると推測できる．解決策としては，ブログ収集する際にストップワードを設定しておき，ストップワードが出現するブログは事前除去するなどが考えられる．後者は，動詞の断片化と同様に，「小向東芝町」「20 日」などの名詞が分割され，その断片として出力されたのが原因である．解決策としては，同様に，複合語処理によりキーワードとして適切な塊で抽出するというのが考えられる．

表 4 動詞分類結果（行動属性抽出）

| 分類 | 出現頻度（延べ数） | 割合（%） |
|---|---|---|
| 行動動詞 | 768,526 | 82.9 |
| 非行動動詞 | 66,475 | 7.2 |
| 未分類動詞 | 92,055 | 9.9 |

表 5 出現頻度上位 5 位以内の動詞（行動属性抽出，what）

| 名詞 | 動詞 | 分類 | 出現頻度（延べ数） |
|---|---|---|---|
| 物件 | 探す | 行動動詞 | 1,738 |
| 気 | つける | 未分類動詞 | 1,340 |
| 写真 | 撮る | 行動動詞 | 1,072 |
| 番号 | 伝える | 行動動詞 | 924 |
| 声 | かける | 未分類動詞 | 830 |

表 6 出現頻度上位 5 位以内の動詞（行動属性抽出，where）

| 名詞 | 動詞 | 分類 | 出現頻度（延べ数） |
|---|---|---|---|
| 町 | 探す | 行動動詞 | 179 |
| 寺 | 探す | 行動動詞 | 128 |
| 物件 | 探す | 行動動詞 | 121 |
| コンビニ | 買う | 行動動詞 | 91 |
| スーパー | 買う | 行動動詞 | 84 |

表 7 出現頻度上位 5 位以内の動詞(行動属性抽出, when)

| 名詞 | 動詞 | 分類 | 出現頻度(延べ数) |
|---|---|---|---|
| 夏 | 休む | 行動動詞 | 2,192 |
| 日 | 見る | 行動動詞 | 1,147 |
| 時間 | かかる | 未分類動詞 | 1,057 |
| 日 | 行く | 行動動詞 | 868 |
| 今日 | 行く | 行動動詞 | 861 |

表 8 出現頻度上位 5 位以内の動詞(行動属性抽出, whom)

| 名詞 | 動詞 | 分類 | 出現頻度(延べ数) |
|---|---|---|---|
| 自分 | 合う | 非行動動詞 | 140 |
| 私 | 言う | 行動動詞 | 123 |
| 私 | 見る | 行動動詞 | 120 |
| 友人 | 話す | 行動動詞 | 113 |
| 業者 | 願う | 行動動詞 | 96 |

表 9 出現頻度上位 5 位以内の動詞(行動属性抽出, how)

| 名詞 | 動詞 | 分類 | 出現頻度(延べ数) |
|---|---|---|---|
| 地図 | 見る | 行動動詞 | 197 |
| テレビ | 見る | 行動動詞 | 189 |
| ネット | 検索する | 行動動詞 | 185 |
| ネット | 調べる | 行動動詞 | 178 |
| 物件 | 探す | 行動動詞 | 121 |

ここまで, 行動動詞判定辞書を利用することで, ブログから行動, および行動属性を一定割合抽出できることが分かった. それぞれの課題を解決した後の, 次のステップとしては, 抽出した行動, および行動属性の体系化が挙げられる. 本提案手法では, 行動間の関係性, 行動と行動属性の間の関係性を一切考慮せずに, それぞれが独立したものとして考えた. ところが,「本, で, 調べる」と「ネット, で, 検索する」は類似の行動,「切符, を, 買う」と「電車, で, 移動する」は連続した行動であるように, それぞれの行動, 行動属性には何かしらの関係性がある場合が多い. そこで, 行動間の関係性, 行動と行動属性の間の関係性をオントロジーや意味ネットワークのようなグラフ形式で体系化しておけば, 複数の行動を一つの行動として適切な粒度にまとめる, また行動の一部が記録されていない場合に前後の行動から補完するといったことが可能となる. そうなれば, テキスト情報からのライフログ自動取得が一層容易となるであろう.

## 5. 関連研究

関連研究として, ブログ, Twitter からのユーザ情報抽出の研究を紹介する.

倉島ら([倉島 10])は, ブログテキストから時間, 空間, 動作, 対象, 感情の 5 要素を抽出する手法を提案する. 時間, 空間は RSS 配信のメタデータから, 動作, 対象は自然言語処理と辞書照合により, 感情は感情語辞書からそれぞれ抽出する. さらに, 動作と対象を行動という 1 つの要素にまとめた上で, 各要素の共起頻度から相関ルールを抽出する. 以上の処理により, 人間の経験情報を構造化, 状況と行動と主観との間の関係を知識化することが可能となった. グェンミンら ([グェンミン 10]) は, 自己教師あり学習を用いて, Twitter API で取得したツイート (Twitter で投稿されたメッセージ) 中に現れる動作とその基本属性(行動主, 対象, 時間, 場所)を自動的に抽出する手法を提案する. 形態素解析結果を系列ラベリングしたものを素性とし, 係り受け関係, 固有表現, 構文リストを利用することで特徴モデルを作成する. 系列ラベリングを利用した自己教師あり学習を用いることで, 低頻度の動作であっても低コストで抽出することが可能となった. Hannon ら ([Hannon 10]) は, Twitter API で取得したユーザのツイート, フォローしているユーザのツイート, およびフォローされているユーザのツイートからプロファイルを作成する, レコメンドシステムを提案した. それぞれ直近 100 件のツイートを対象に, tf-idf によりプロファイルに使用するキーワード候補を決定する. そして, それぞれのキーワード候補を組み合わせることでプロファイルを作成し, コンテンツベースドフィルタリング, 協調フィルタリングによりコンテンツを推薦する. Twitter を利用することで, リアルタイム性の高いレコメンドシステムを実現することが可能となった.

1 つ目の関連研究は, ブログから人間が主体となる動作, およびその対象を抽出しており, 本研究と共通している部分が多い. ところが, 状況と行動と主観との間の関係性を解明することを目的としており, 行動の体系化を目的としている本研究とはスタンスが異なる. 行動を体系化する場合には, 例えば, 行動をさらに観察可能な外面的行動と観察不可能な内面的行動に区別するなど, 行動をそれ自身が持つ意味によって構造化する必要がある. 2 つ目, 3 つ目の関連研究は, Twitter に投稿されたツイートから何かしらのユーザ情報を抽出しているが, ユーザの主体的動作, 内面的行為, および客観的事実を区別していない. 目的によってはこれらを区別する必要はないかもしれないが, ライフログの自動取得を目的とする場合にはこれらを区別することが必須となる.

## 6. まとめ

行動動詞判定辞書を利用した, 行動, および行動属性をブログから抽出する手法を提案した. これにより, ブログや Twitter に代表されるライフストリームからの, ライフログ自動取得の可能性の一端を示すことができた. 課題としては, 砕けた日本語に対しても有効な形態素解析エンジンの利用, 同一の行動であっても漢字と平仮名で表現される場合があるといった表記揺れ問題への対応, 動詞や名詞をキーワードとして適切な塊で抽出するための複合語処理の追加などが挙げられる.

今後は, 行動, および行動属性の体系化を行い, テキスト情報からのライフログ自動取得をさらに進めていく.

### 参考文献

[Hannon 10] J. Hannon, M. Bennet. and B. Smyth: Recommending Twitter Users to Follow using Content and Collaborative Filtering Approaches, RecSys 2010, pp. 199-206, 2010

[Kwak 10] H. Kwak, C. Lee, H. Park and S. Moon: What is Twitter, a Social Network or a News Media?, WWW2010, pp. 591-600, 2010

[Sellen 07] A. J. Sellen, A. Fogg, M. Aitken, S. Hodges, C. Rother and K. Wood: Do life-logging technologies support memory for the past? An experimental study using sensecam, CHI 2007, pp. 81-90, 2007

[倉島 08] 倉島健, 藤村考, 奥田英範: 大規模テキストからの経験マイニング, 電子情報通信学会 第 19 回データ工学ワークショップ, A1-4, 2008.

[グェンミン 10] グェンミンティ, 川村隆浩, 田原康之, 大須賀昭彦: Twitter からの人間行動属性の抽出, 信学技報 AI2010-4, pp. 19-23, 2010.

[竹内 08] 竹内孔一, 乾健太郎, 竹内奈央, 藤田篤: 意味の包含関係に基づく動詞項構造の細分類, 言語処理学会 第 14 回年次大会, B5-2 , 2008.