特集論文 「エージェント」

条件付確率場と自己教師あり学習を用いた行動属性の自動抽出と評価

Automatic Extraction and Evaluation of Human Activity Using Conditional Random Fields and Self-Supervised Learning

グェン ミンティ Nguyen Minh The
電気通信大学大学院情報システム学研究科
Graduate School of Information Systems, The University of Electro-Communications
minh@ohsuga.is.uec.ac.jp

川村 隆浩 Kawamura Takahiro
（同 上）
kawamura@ohsuga.is.uec.ac.jp, http://www.ohsuga.is.uec.ac.jp/~kawamura/

中川 博之 Nakagawa Hiroyuki
（同 上）
nakagawa@ohsuga.is.uec.ac.jp, http://www.is.uec.ac.jp/staff/217.html

田原 康之 Tahara Yasuyuki
（同 上）
tahara@ohsuga.is.uec.ac.jp, http://www.is.uec.ac.jp/staff/218.html

大須賀 昭彦 Ohsuga Akihiko
（同 上）
akihiko@ohsuga.is.uec.ac.jp, http://www.ohsuga.is.uec.ac.jp/~ohsuga/

keywords: Real-world Agent, Human Activity, Web Mining, Semantic Network, Self-Supervised Learning

Summary

In our definition, human activity can be expressed by five basic attributes: actor, action, object, time and location. The goal of this paper is describe a method to automatically extract all of the basic attributes and the transition between activities derived from sentences in Japanese web pages. However, previous work had some limitations, such as high setup costs, inability to extract all attributes, limitation on the types of sentences that can be handled, and insufficient consideration interdependency among attributes. To resolve these problems, this paper proposes a novel approach that uses conditional random fields and self-supervised learning. Given a small corpus sample as input, it automatically makes its own training data and a feature model. Based on the feature model, it automatically extracts all of the attributes and the transition between the activities in each sentence retrieved from the Web corpus. This approach treats activity extraction as a sequence labeling problem, and has advantages such as domain-independence, scalability, and does not require any human input. Since it is unnecessary to fix the number of elements in a tuple, this approach can extract all of the basic attributes and the transition between activities by making only a single pass. Additionally, by converting to simpler sentences, the approach can deal with complex sentences retrieved from the Web. In an experiment, this approach achieves high precision (activity: 88.9%, attributes: over 90%, transition: 87.5%).

1. はじめに

計算機がユーザの行動意図を把握し，それに応じたサービスを提供することは，ユビキタスコンピューティング [Poslad 09] とソーシャルコンピューティング [Ozok 09, NikkeiBP] の双方において重要な課題とされている．例えば，各消費者の行動に基づいて，広告を配信するサービス（One to One マーケティング [Peppers 99]）や図1に示すようなユーザの経験共有サービスなど最適な商品・行動パターンを提供することが考えられる．本研究の最終的な目的は，このようなユーザの行動意図に応じたサービスを提供する実世界指向エージェントを実現することである．

得られた情報リソースから行動意図を認識するためには，行動の構成要素（行動属性）の把握と行動間の関係（遷移，因果関係など）が必要である．これを予め定義しておくことは膨大なコストがかかるだけでなく，未知の意図にも対応できず問題がある．一方，携帯電話やRFID，各種センサーを利用し，イベントの記録を行動履歴として，そこから頻出するイベント連鎖を発見する研究 [NTTDocomo, KDDI] がある．しかし，川村ら [川村 08] が指摘するように，このアプローチでは，

- イベントデータにはノイズが多く，意味のあるイベントの連鎖を見つけるのは難しい．
- 個人の履歴からだけでは，その人にとって意外性のある発見ができない．他人の履歴を用いるには処理

データ量やセキュリティ，プライバシーなどに問題がある．

- 行動パターンを発見し，手動でルール化するには，構築，保守に多大なコストが必要となる．

といった多くの問題がある．このため，本研究では，イベントデータからではなく，Web コーパスから行動を表す文を収集して行動属性と行動間の関係を抽出する．しかし，Web コーパスから行動属性と行動間の関係を抽出する先行研究では，抽出のための準備コストが大きいこと [川村 08] や，抽出できる行動属性が少ないこと [Perkowitz 04, 川村 08]，適用可能な文の種類が少ないこと [Perkowitz 04, 倉島 09]，行動属性間の係り受け関係を十分に考慮されていないこと [Perkowitz 04, 倉島 09] などといった問題がある．

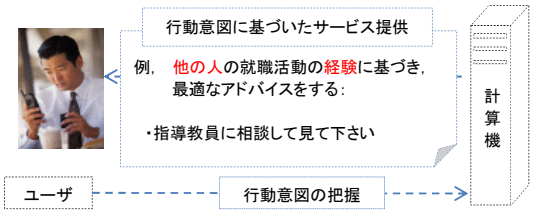

図 1 経験共有サービス：行動意図を把握して，他の人の経験に基づき最適なアドバイスをする

そこで本論文では，日本語の Web ページを対象とし，文に現れる基本行動属性（行動主，動作，対象，場所，時間）と行動間の遷移の自動抽出手法を提案する．提案手法は Web コーパスからのバイナリリレーション抽出の最先端技術である O-CRF[Banko 08] をベースとし，条件付確率場（Conditional Random Fields）と自己教師あり学習（Self-Supervised Learning）*1を利用する．まず，少量のサンプルデータ（小規模なコーパス）から各文に現れる行動属性と行動間の遷移を抽出し，訓練データを自動的に作成する．次に，条件付確率場を用いて訓練データの特徴（行動属性の特徴，行動属性間の係り受け関係，行動間の遷移）を学習し，特徴モデルを作成する．最後に，この特徴モデルを用いて未知データ（Web コーパスの文）の行動属性と行動間の遷移を抽出する．

本論文は，次のような構成をとる．第 2 章では，条件付確率場を説明する．第 3 章では，条件付確率場と自己教師あり学習を用いた行動属性と行動間の遷移の自動抽出手法を提案する．第 4 章では，評価実験，実験結果の考察，提案手法の有効性を述べる．第 5 章では，関連研究を解説し，本論文が提案する手法との比較を行う．最後に第 6 章では，今後の課題と合わせて本論文をまとめる．

*1 サンプルデータから訓練データを自動的に作成するので，Self-Supervised Learning と呼ぶ．

2. 条件付確率場とは

条件付確率場（Conditional Random Fields）とは，John D. Lafferty ら [Lafferty 01] が提案した系列ラベリング問題に適用するための識別モデル（discriminative model）である．

入力データを x（例えば，文），出力データを y（例えば，固有名詞）とするとき，学習モデルに求められる性質は，x が与えられたときに対応する y が正しく出力されるということである．条件付確率場 (以下 CRF) は一つの指数分布モデルで，出力系列 $y = y_1, y_2, ..., y_n$ の入力列 $x = x_1, x_2, ..., x_n$ に対する条件付確率 $P(y|x)$ を表す．

$$P(y|x) = \frac{\exp\langle\alpha, \Phi(x,y)\rangle}{\Sigma_{y' \in O(x)} \exp\langle\alpha, \Phi(x,y')\rangle} \quad (1)$$

但し，$\Phi(x,y)$ は系列 $y = y_1, y_2, ..., y_n$ 上のパスの全ての特徴ベクトルを足し合わせたものであり，α はモデルのパラメータである．$O(x)$ は x 上に定義される全ての系列集合である．そして CRF において，新たな x が与えられたときの出力予測 $\hat{y}$ は

$$\hat{y} = argmax_{y \in Y} P(y|x) \quad (2)$$

となる．この出力は，Viterbi アルゴリズム [Forney 73] を用いることで効率良く解くことができる．

CRF は識別モデルであり，重複する特徴をモデルに組み込むことができる．通常の識別モデルとの違いは，出力が出力集合の部分集合ではなく，系列となる点である． CRF は文字種や周辺の単語情報などといった素性を素性関数という形で柔軟に投入でき [Kudo 04], Hidden Markov Models や Maximum Entropy Markov Models の問題点（label bias，length bias）を自然にかつ有効に解決できる．そして CRF は，品詞付与 [Lafferty 01]，テキストチャンキング [Sha 03]，固有表現抽出 [McCallum 03]，形態素解析 [Kudo 04] などといった系列ラベリング問題に適用され，いずれにおいても高い精度を示している．

3. 条件付確率場と自己教師あり学習を用いた行動属性と行動間の遷移の自動抽出

3·1 行動属性の定義

行動の核となる要素は「行動主」,「動作」と「対象」である．そして，ユーザの状況に応じた最適な情報を提供するために，“ どこで ”,“ どんな時に ”,“ いつ ”動作が行われるかは重要である．このため，本論文では，人間の行動は「行動主」,「動作」,「対象」,「場所」,「時間」という５つの基本属性から成ると定義している．そして，これらの属性に以下のようなラベル（Who，Action，What，Where，When）を付ける．

(1) 行動主：Who
(2) 動作：Action
(3) 対象：What

（4） 場所：Where
（5） 時刻（いつ），場面（どんな時に）：When

行動間の遷移ラベルは Next（次）又は After（後）と設定する．本論文の課題は，日本語 Web ページの文中に現れる行動の基本属性と行動間の遷移を自動的に抽出することである．例えば，“ ご飯を食べる前に、太郎は手を洗います ”という文に対して，属性と行動間の遷移を抽出し，グラフで表現すると図 2 のようになる．

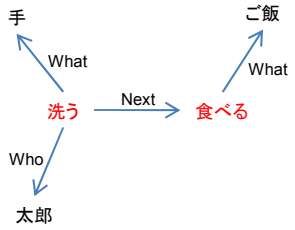

図 2 “ ご飯を食べる前に、太郎は手を洗います ”に現れる行動の基本属性と行動間の遷移

「動作」は行動の中核であるため，この属性がない場合，行動として扱っていない．そして，行動を表す文では，「動作」の単独ではなく，「動作」と 1 つ以上の他の属性（「行動主」，「対象」，「場所」，「時間」）を含む必要がある．そのため，本論文で扱う「行動を表す文」というのは，「動詞句と名詞句を持つ文」（例えば，秋葉原へ行く）又は「名詞句と名詞句が助詞“ を ”で結ばれた文」（例えば，英語を勉強）となる．そして，“ いる ”と“ ある ”は存在を表す動詞なので，対象外としている．

3·2 行動属性抽出の難しさ

日本語 Web ページの文中に現れる行動属性と行動間の遷移の抽出については，以下に示すような難しさがある．

（1） CRF を適用した先行研究の多くは，単語の境界位置が明確であることを想定している．しかし，日本語はスペース文字がなく，明示的な単語境界がない言語である．このため，日本語 Web ページの文において，CRF を適用するために，分かち書きを行う必要がある．したがって，分かち書きの出力が誤ると，属性の判定も誤ってしまう可能性が高い．
（2） 英語における文の構成の大部分は，Subject-Verb-Object である [Banko 08]．しかし日本語では，文の構成は自由度が高く，様々なタイプがある．
（3） Web コーパスは非常に膨大であり多様性を持つ．
（4） ブログや SNS といった CGM（Consumer Generated Media）では，複雑かつ文法的に正しくない文が多い．そして CGM から取得する文では，顔文字や“ えーっと ”，“ 。。。 ”などのようノイズ文字列を含まれる文が多い．
（5） 文によって現れる行動属性の数と位置が変わるので，行動属性にマッチする正規表現パターンの作成は非常に難しい．

3·3 提案手法のアーキテクチャ

Web コーパスは膨大であり，かつ多様性を持つ．このため，提案手法は機械学習アプローチを適用して Web コーパスから取得した文中に現れる行動属性と行動間の遷移を抽出する．また，訓練データを人手で作成すると膨大なコストがかかるため，自動的に作成する．図 3 に示すように，提案手法のアーキテクチャは訓練データの自動作成モジュール（図 3 の I）と行動属性の自動抽出モジュール（図 3 の II）という 2 つのモジュールに分割する．

訓練データの自動作成モジュールは人手でラベル編集，初期インスタンスの作成，行動のドメインの定義などの必要がなく，訓練データを自動的に作成できる．まず，Wikipedia の人物カテゴリ [Wikipediaa] から少量の文書を取得して，サンプルデータとして利用する．次に，サンプルデータの前処理を行い，各文に現れる行動属性と行動間の遷移を抽出する．そして，訓練データと特徴モデルを自動的に作成する．

行動属性の自動抽出モジュールは，まず Web コーパスから取得した文書の前処理を行う．この前処理では，行動を表さない文を削除して，行動を表す文中にあるノイズ文字列（…，えーっと，顔文字など）を削除する．次に，行動を表す文を単純化してテストデータを作成する．最後に，訓練データの自動作成モジュールで作成された特徴モデルを用い，テストデータの行動属性と行動間の遷移を自動的に抽出する．

各モジュールの詳細を以下に示す．

§1 訓練データの自動生成モジュール

以下，“ キムチを食べた後，太郎は歯を磨く ”という例文を用いて，訓練データの自動作成法を説明する．

（1） 文字コードの変換，文書から行動を表す文の取得などサンプルデータの前処理を行う（図 3 の I.1）．
（2） 図 4 に示すように Mecab で形態素解析を行い，文の単語と単語の品詞番号を取得する（図 3 の I.2）．

キムチ	38
を	13
食べ	31
た	25
後	66
、	9
太郎	44
は	16
歯	38
を	13
磨く	31

図 4 例文の単語と単語の品詞番号

（3） Cabocha[Kudo 02] で係り受け解析を行い，NP（Noun Phrase）と VP（Verb Phrase）の係り受け関係を把握

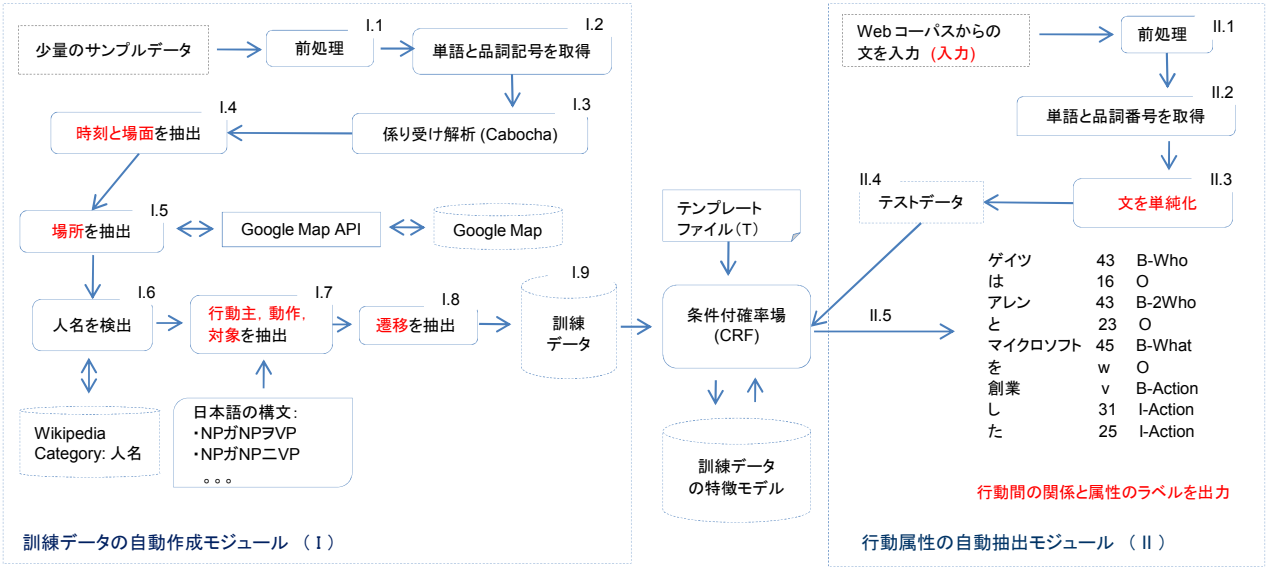

図 3 提案手法のアーキテクチャ

する（図 3 の I.3）．図 5 に示すように，形態素解析と係り受け解析をした後，“ キムチを食べた後，太郎は歯を磨く ”の構成は「NP ヲ VP-タ後、NP ハ NP ヲ VP」であることが分かる．また，この文は二つの行動を含み，行動 B が起こった後に行動 A が起こることを表す．

(4) 係り受け解析の結果を用いて，文の VP と係り受け関係をもつ時間表現（時刻，場面）を抽出する（図 3 の I.4）．“ キムチを食べた後 ”は“ VP-タ後 ”という正規表現パターンにマッチするので，場面を表す節として認識される．この節にラベルを付けると図 6 のようになる．但し，What は行動の対象のラベル，Action は動作のラベル，After は“ 後 ”のラベル，2 は 2 番目の行動，B は Begin（属性の先頭），I は Inside（属性の途中），O は Other（属性以外）という意味を表す．この BIO は自然言語処理のテキストチャンキング問題 [CoNLL] に良く使われる記号である．

(5) 係り受け解析結果を用いて，文の VP と係り受け関係をもつ場所を抽出する（図 3 の I.5）．

(6) 長い人名をカタカナで書くと，Mecab の解析精度が落ちる．この問題を解決するために，Mecab の解析結果に加えて，Wikipedia の人名カテゴリ [Wikipediab] を活用して文の人名を検出する（図 3 の I.6）．この方式以外にも，人名カテゴリ中の人名を Mecab の辞書に組み入れるという方式も考えられる．しかし，Mecab の辞書に組み入れる方式より，Wikipedia を直接に問い合わせる方式の方が柔軟性は高い．

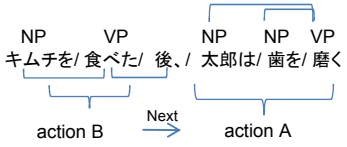

図 5 例文の形態素解析と係り受け解析の結果

キムチ	B-2What
を	O
食べ	B-2Action
た	I-2Action
後	B-After
、	O

図 6 例文の行動 B

(7) 係り受け解析結果を用いて，VP と係り受け関係がある NP を抽出する．次に，日本語の構文リスト（NP ガ NP ヲ VP，NP ガ NP ニ VP など）を用いて，これらの NP はどれが行動主，対象であるかを判定する（図 3 の I.7）．“ 太郎は歯を磨く ”にラベルを付けると図 7 のようになる．但し，Who は行動主のラベルである．

太郎	B-Who
は	O
歯	B-What
を	O
磨く	B-Action

図 7 例文の行動 A

(8) VP-タ後，VP-前に，VP-タラ，次に，そしてなどのような日本語の文法パターンを採用して，行動間の遷移を抽出する（図 3 の I.8）．

(9) 以上の解析結果と単語の品詞記号を合わせて，訓練データを作成する（図 3 の I.9）．例文の訓練データは図 8 のようなものである．

キムチ	38	B-2What
を	w	O
食べ	v2	B-2Action
た	25	I-2Action
後	66	B-After
、	9	O
太郎	44	B-Who
は	16	O
歯	38	B-What
を	w	O
磨く	v	B-Action

図 8 例文の訓練データ

次に，行動の基本属性（行動主，動作，対象）の抽出と「場所，時刻，場面」の判定について述べる．

- **行動主，動作，対象の抽出**

提案手法は下記のような日本語の構文パターンを考慮し｛行動主，動作，対象｝を抽出する．

(1) “ ヲ ”,“ 二 ”又は“ へ ”という助詞がある文
- ○{O，C}，{ ヲ，二，へ }，V （例：映画を見る）*2
- ○S，{O，C}，{ ヲ，二，へ }，V （例：太郎は英語を学んでいる）
- ○{O，C}，{ ヲ，二，へ }，V，S（例：マイクロソフトを創業したのはゲイツです）

(2) “ ガ ”と“ ハ ”を含む文
- ○S ガ V ハ {O，C}（例：太郎が見た映画は面白い）
- ○S ガ V{C} ハ {O}（例：ゲイツが創業した会社はマイクロソフトです）
- ○S ハ N ガ V（例：太郎はラーメンが大好きなのでよく食べる）

(3) その他
- ○ヲ N （例：論文を作成）
- ○N ガ（ハ）V（例：太郎が読む）
- ○N ヲ N 二（例：風景を写真に撮りました）

行動主，動作，対象はそれぞれ S，V，O に当たる．

- **場所の判定**

抽出精度を向上するために，形態素解析結果に加えて，Google Maps API[Google] を用いて場所を判定する．入力名詞句に対して，Google Maps API のレスポンスがアドレスであれば，この入力名詞句は場所であると判定する．

- **時刻 (いつ)**

今，月曜日，上旬，時などのような文字列を含む時間表現（約 100 表現）を時刻として判定する．

- **場面（どんな時に）**

V-た後，V-たとたん，V-前になどのような日本語の文法パターンを採用して，場面を判定する．

*2 以下において，S は Subject（主語），O は Object（述語），C は Complement(補語)，V は Verb（動詞），N は Noun(名詞) という意味を表す．

§2 行動属性の自動抽出モジュール

行動属性の自動抽出モジュールの主なタスクは以下のようなものである．

(1) Web から取得した文書の前処理を行い,「行動を表す文」を取得する．（図 3 の II.1）．

(2) 図 9 に示すように HTML タグの解析結果に加えて，形態素解析を行い，単語とその品詞番号を取得する（図 3 の II.2）．

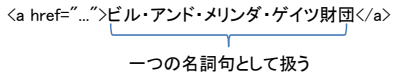

図 9 HTML タグを解析して名詞句を取得

(3) 形態素解析結果に基づいて，文の名詞句と動詞句を把握し単純記号（NP，VP）に置き換える．但し，変換された名詞句と動詞句の品詞番号を保持する．これを行うことによって，文を単純化でき，テストする時にラベル推定のエラーを防止できる．文を単純化してテストデータを作成すると図 10 のようになる（図 3 の II.3 と II.4）．

(4) 条件付確率と訓練データの特徴モデルを用い，テストデータの行動属性と行動間の遷移を自動的に抽出する（図 3 の II.5）．

図 10 文の単純化：名詞句と動詞句を単純化

3·4 条件付確率場を用いた行動属性の抽出

提案手法では，正規表現パターンではなく系列ラベリングを用いて行動属性を抽出する．例えば,“ 太郎は 24 日に修士論文を提出する ”という文を系列ラベリングで表すと図 11 のようになる． 2 章で述べたように系列ラベリング問題に適用するための学習モデルのうち、条件付確率場が高い精度を得られるので，本手法はこれを利用する．そして，データ（訓練データ，テストデータ）フォー

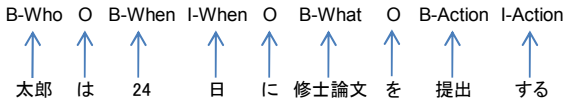

図 11 行動属性を系列ラベリングで表現

マットとテンプレートファイルの設計方針は次の 2 つがある．一つ目は，訓練データの特徴を全て吸収できることである．もう一つは，学習とテストの計算量を最小限にすることである．これらの設計方針に従い，本手法は以下のようなデータフォーマットとテンプレートファイルを設計する．

単語	品詞記号	ラベル
太郎	44	B-Who
は	16	O
北海道	46	B-Where
で	d	O
白い	10	B-What
恋人	38	I-What
を	w	O
買っ	v	B-Action
た	25	I-Action

図 12 訓練データのフォーマット

• 訓練データのフォーマット

図 12 に示すように，訓練データは単語列，品詞記号列，ラベル列から構成される．単語列は文の単語を表す．品詞記号例は Mecab の品詞番号，助詞の記号 (に→ n，を→ w，へ→ h，で→ d)，動詞の記号 (v) を表す．ラベル列は行動属性と行動間の遷移のラベルを表す．

• テストデータのフォーマット

図 13 に示すように，テストデータは単語列と品詞記号列だけがある．ラベルはテストする際にシステムが推測して自動的に付与する．

単語	品詞記号
太郎	44
は	16
ラーメン	38
を	w
食べる	v

図 13 テストデータのフォーマット

• テンプレートファイル

テンプレートファイル（図 3 の T）とはデータの特徴を表すためのファイルである．提案手法が利用する素性（特徴）は単語，品詞，助詞である．テンプレートファイルはこれらの素性と係り受け関係を扱う．そして，長い文に対応させるために，サイズ 7 のウィンドウ（%x[-3,*] ～%x[3,*]）を採用する．図 14 にテンプレートファイルの全体を示す．但し，%x[i,j] は現在の位置からの相対位置で i 行目の j 番目の列の要素を指す．また，U**はテンプレートの記号である．

```
# POS column
U10:%x[-3,1]
U11:%x[-2,1]
U12:%x[-1,1]
U13:%x[0,1]
U14:%x[1,1]
U15:%x[2,1]
U16:%x[3,1]
U17:%x[-3,1]/%x[-2,1]
U18:%x[-2,1]/%x[-1,1]
U19:%x[-1,0]/%x[0,1]
U110:%x[0,1]/%x[1,1]
U112:%x[1,1]/%x[2,1]
U113:%x[2,1]/%x[3,1]

U25:%x[1,1]/%x[2,1]/%x[3,1]
```

```
# Word column
U00:%x[-3,0]
U01:%x[-2,0]
U02:%x[-1,0]
U03:%x[0,0]
U04:%x[1,0]
U05:%x[2,0]
U06:%x[3,0]
U07:%x[-1,0]/%x[0,0]
U08:%x[0,0]/%x[1,0]

#POS column 's junction
U21:%x[-3,1]/%x[-2,1]/%x[-1,1]
U22:%x[-2,1]/%x[-1,1]/%x[0,1]
U23:%x[-1,1]/%x[0,1]/%x[1,1]
U24:%x[0,1]/%x[1,1]/%x[2,1]
```

図 14 テンプレートファイル

4. 評価実験

4·1 前処理

3·3 節に説明したように，本論文では Web コーパスから取得した文章そのものではなく，前処理を行って，3·1 節に定義した「行動を表す文」を対象に属性と行動間を抽出する．Mecab の解析精度は 100%ではないので，前処理の過程で「行動を表す文」が削除されてしまう可能性がある．また，行動を表さない文が行動を表す文として扱われてしまう可能性もある．一般の文では，Mecab での品詞判定精度は約 98%[Kudo 04] と言われているため，前処理の課程で，誤って捨てられた又は含められた行動を表す文は約± 2%程度だと考えられる．これを確かめるために，我々は Mecab の精度と本論文での「行動を表す文」の定義を含めて，実際の誤差を把握するため，以下の評価実験を行った．

図 15 では，以下のような記号を用いて，前処理について説明する．

→ A は，Web コーパスから取得した文の集合である．

→ B は，A の中にある行動を表す文の集合である（行動を表す文の正解集合）．

→ C は，A を対象に前処理を行った結果として取得され，行動属性と行動間の遷移を抽出する対象文の集合である．

評価実験のデータセットを【付録 D】に示す．これは「秋葉原」という検索キーワードで twitter[Twitter] から検索した 105 文である．105 文中，75 文が行動を表す文（集合 B）であった（人手により確認）．前処理過程では 33 文（33/105=31.43%）が削除され，72 文が集合 C となった．集合 C の中に，行動を表す文は 70 文（70/75=93.33%）

があり，行動を表さない文は 2 文があった．捨てられた文の中に，行動を表す文は 5 文（5/75=6.67%）であった．よって，前処理過程で捨てられた文は 31.43%で，このうち誤って捨てられた行動を表す文は 6.67%である．行動文によって属性の種類と数が変わるが，約 6.67%の属性を誤って捨ててしまっていることが分かった．

尚，前処理過程で，誤って捨てられた行動を表す文は図 16 の通りである．これらの文では，動詞が“ お客様訪問終了 ”や“ 妹の電子辞書購入 ”などの名詞となっている．また，人は“ 群馬県太田市へ〜。”を行くという行動だと判断できるが，“ 行く ”という動詞が省略されているため，計算機は判断できなかった．

前処理過程で，行動を表さないが行動を表すとして扱われてしまった文は図 17 の通りである．これらの原因は，Mecab が品詞を誤って判断したためである．“ 田中すだれ店。”の文では，Mecab が“ だれ ”は動詞（“ だれる ”）であると判断してしまった．“ でも喫茶じゃないでふ。”の文では，Mecab が“ ふ ”は動詞（“ ふる ”）であると判断してしまった．

上記の前処理の誤りについては，今後，改善を検討していく．

4·2 「行動属性」と「行動間の遷移」の抽出

提案手法を用いることで，先行研究の問題点と 3·2 節に述べた課題を解決できたかどうかに加えて，行動の基本属性と行動間の遷移の抽出精度を明らかにするために，我々は以下の評価実験を行った．

【付録 A】に示すように，評価実験のデータセットは「行動を表す」260 文である．これは Web コーパスからランダムに取得した文の前処理を行った結果である．図 18 は実験データの一部である．

文に現れる全ての行動の基本属性，属性間の係り受け関係，行動間の遷移を正しく抽出することができた場合

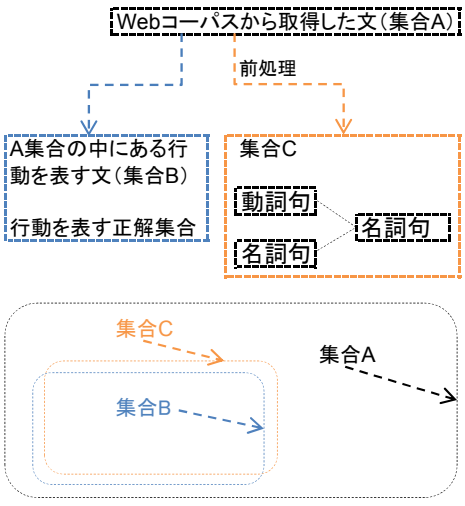

図 15 「A 集合」と「B 集合」と「C 集合」の関係

秋葉原にてお客様訪問終了。
今日はこれから秋葉原で待ち合わせ。
秋葉原で妹の電子辞書購入。
秋葉原ノードでリハ。
これから群馬県太田市へ〜。

図 16 誤って捨てられた行動を表す文

田中すだれ店。
でも喫茶じゃないでふ。

図 17 行動文として扱われてしまった文

に行動の抽出が正解であると判定する．行動の抽出精度は，すべての行動文（260）に対して，行動が正しく抽出された文の割合である．そして，属性の抽出精度は，抽出すべきの属性に対して，正しく抽出された属性の割合である．抽出精度の実験結果（付録 B）は表 1 の通りである．

表 1 抽出精度の実験結果

	抽出すべき対象	正解	精度（%）
行動文	260	231	88.85%
行動主/Who	131	124	94.66%
動作/Action	352	342	97.16%
対象/What	244	234	95.90%
場所/Where	78	71	91.03%
時刻，場面/When	97	95	97.94%
行動間の遷移	16	14	87.50%

実験結果により，本手法では，日本語 Web ページを対象にして，一回の実行（テスト）で文に現れる全ての属性（行動主，動作，対象，場所，時間）と行動間の遷移を自動的に抽出でき，高い抽出精度を得た．しかし，Wikipedia の書き方がある程度の統一性を持つため，現時点の訓練データはまだ「複雑な文」や「正しい文法で記述されていない文」を上手く対応していない．これらの文に対応するために，多様性を持つ訓練データを追加する必要がある．実験結果を通じて以下のようなことが分かった．

- Web コーパスから収集した文の単純化により，ラベル予測のエラーを防止できる．
- 動作は文の動詞句に当たる．本研究が利用する形態素解析ツール Mecab は日本語の動詞を良く認識できる．また，テストデータを作成する前に，動詞句を単純化する．このため，動作の抽出精度は高くなる (97.16%).
- 日本語の場合，動作と対象の間に助詞（ヲ，ニ，ヘ）が現れるケースが多いという特徴をもち，そのことが対象の抽出精度を高めた（95.90%）と考えられる．
- 実験データでは，行動主は人名である文が多い．また，日本語の場合，主語は“ ハ，ガ ”との関係が強いという特徴をもつ．このことが，行動主の抽出精

そして，ゲイツはアレンと共にマイクロソフト社を創業した。
私は毎朝自転車で学校に行きます。
部屋で数学の問題を考えていました。
頭の中で何かがひらめいた。
午前7時、そろそろ他の家族も起き出してきました。
2週間前に、日本へ来たばかりです。
彼は国へ帰る前に、先生のところへお礼に行きました。
日本へ留学する前に、なるべく日本についての本を読んできてください。
ホテルのイリマ・カフェで朝食を済ませてから、部屋へ戻って外出の準備をします。
カウンターで、日本の運転免許証と国際運転免許証を提示して手続きを行います。

図 18 実験データの一部

度を高めた（94.66%）と考えられる．

- GoogleMap API を利用することで複雑なアドレスを扱うことができ，場所の抽出精度を高めた（91.03%）．
- 本手法では Mecab での解析結果に加えて，日本語の時間表現（時刻，場面）に対応したので，時刻と場面の抽出精度は高くなった（97.94%）．

そして，本手法では，1 台の PC（CPU：3.2Ghz，RAM：3.5GB）を用いた場合に，260 文の抽出時間は約 0.11 秒であったが，一方で，Cabocha を用いて，この 260 文の係り受けのみを解析した場合の処理時間は約 19 秒であり，提案手法の 172 倍であった．従って，提案手法は Web コーパスのような大規模コーパスに対して有効な手法であると言える．

第 1 章に示した先行研究の問題点に対して，提案手法は以下の対策を行う．

- 行動のドメインを定義せず，訓練データを自動的に作成することで，準備コストがかからない．
- CRF を系列ラベリングに適用し，更にタプルの要素数を固定しないため，一回のテストで文中に現れる全ての行動属性を抽出できる．
- 機械学習の適用と Web コーパスから取得した文の単純化により，提案手法は単文や複文など様々な文に対応できる．
- 訓練データが属性間の係り受け関係を含むため，テストする時に文中に現れる行動属性間の係り受け関係を推定できる．
- 公開されている Web ページから行動データを取得するので，プライバシー問題を回避できる．

また，3·2 節に示した日本語 Web ページにおける行動属性抽出の難しさに対して，本論文は以下の対策を行う．

(1) 形態素解析を行い単語の境界位置を把握することで，CRF を適用することができる．
(2) 日本語の文には様々なタイプがあるので，提案手法は表 2 に示す構文に対応する訓練データを作成する．
(3) 機械学習の適用と文の単純化により，膨大かつ多様性をもつ Web コーパスに適用できる．
(4) 前処理を行って，文中にあるノイズ文字列を削除する．そして，複雑かつ正しい文法で記述されていない文に対応するために，文の単純化と訓練データの拡張を行う．
(5) 文中に現れる行動属性にマッチする正規表現パターンの作成が困難である問題に対して，提案手法はパターンマッチングではなく系列ラベリングを利用する．

4·3 O-CRF との比較

表 2 O-CRF と本手法との比較

	O-CRF	提案手法
Web ページの言語	英語	日本語
抽出対象	リレーション	人間の行動
タプルの要素の数	3	2〜6
対応可能な構文	S-V-O	{O，C}，V S，{O，C}，V {O，C}，V，S S ガ V ハ {O,C} S ガ VC ハ O S ハ N ガ V ヲ N N ガ（ハ）V N ヲ N 二
リレーションはエンティティの間に出現する必要がある	YES	NO
エンティティを事前に判定しておく必要がある	YES	NO
エンティティを意味的に分類する	NO	YES
適合率（精度）	86.6%	88.85%
再現率	45.2%	82.18% (88.85-6.67)

O-CRF[Banko 08] は自己教師あり学習と CRF を用いて，英語 Web ページの文に現れるバイナリリレーションを抽出する．図 19 に示すように，リレーションはエンティティの間に現れる必要がある【付録 C】．そして，表 2 に示すように，O-CRF では，エンティティを事前に判定しておくので，ラベルの推定はリレーションだけである．一方，提案手法は，一つのラベルだけではなく，文に現れるすべての行動属性と行動間の遷移のラベルを推定する必要がある．また，日本語の文を対象するので，文によって属性の数と位置は変わる．このため，提案手法では，エンティティの事前判定，又はリレーションがエンティティの間に現れるといった設定ができない．つまり，O-CRF の解決課題よりも提案手法が解決する課題が困難であると考えられる．

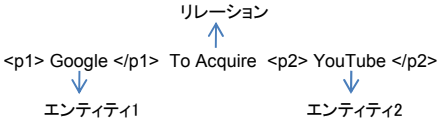

図 19 O-CRF の実験データ：リレーションはエンティティの間に現れる必要がある

提案手法は全ての前処理を行った後の文（前述の集合 C）に対して，文中に現れる行動属性（行動主，動作，対象，場所，時間）と行動間の遷移（次，後）を抽出する．4·1 節に示したように，前処理過程で約 6.67%の「行動を表す文」を誤って捨ててしまったので，提案手法の再現率は精度（適合率）より約 6.67 %低い値となり，82.18% (88.85-6.67) である．

Web ページの言語と抽出対象が異なるため，直接比較することができないが，本手法が O-CRF より高い精度を得たのは以下のようなことによる．

- 属性の数と位置を制限する必要がなく，抽出すべき属性を全て抽出する．
- 全ての標準構文に対応する訓練データを作成する．
- テストを行う前に，Web コーパスから取得した文を単純化する．
- 動作と対象の間に助詞（ヲ，ニ，ヘ）が現れる日本語の特徴に対応する．
- テンプレートファイルのウィンドウサイズが大きいので，長い文にも対応可能である．
- 時間，場所，人名の抽出を工夫する．

4·4 提案手法の有効性

本研究は，最終的には行動推薦・行動ターゲティングを目指しており，ユーザの現在の行動と次の行動又は行動の原因を把握できれば，ユーザの状況に応じた最適な情報（広告，商品，店舗，行動パターンなど）を提供できると考えている．現在，多くの企業や研究所で，ユーザの行動モデルや行動推測などの研究・開発が行われ，携帯端末向けのサービスなどが試行されている．例えば，NTT Docomo の My Life Assist Service[NTTDocomo] では，ユーザの行き先を予測し，行き先周辺の店舗情報を提供する．KDDI 研究所の「ケータイ de ライフログ」[KDDI] では，ユーザのライフログ（いつ・どこで・誰と・何をしたか，何に興味をもったのか）を収集・管理し，行動を解析し，適切な情報を提供することを目指している．

本研究は，上記のアプリケーション等に加えて，コンテキストアウェアコンピューティング [Matsuo 07] やユビキタスコンピューティング [Poslad 09]，ソーシャルコンピューティング [Ozok 09] など多くの分野に適用できると考えている．現在，我々は街歩きに着目し，Web コーパス（weblogs，twitter 等）から“ 秋葉原 ”に関するユーザの行動と行動間の関係を抽出して，意味ネットワークを構築することを試みている．図 20 に示すように，この意味ネットワークでは，ノードは動作，ノードの周辺は行動の属性（行動主，対象，時間，場所），ノード間のリンクは行動間の関係を表す．この意味ネットワークを参照して，ユーザの行動データから行動意図を推測し，経験に基づく賢い行動パターンを推薦することを目指している．

上記の目的を実現するためには，行動と行動間の関係

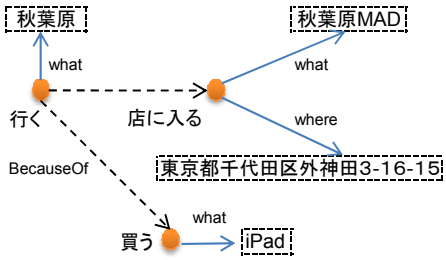

図 20 秋葉原に関する行動を表す意味ネットワーク（一部）

ができるだけ多く意味ネットワークに含まれている必要がある．つまり，意味ネットワークのノードとリンクをどの程度構築できたかが，提案手法の有効性を評価する指標となる．提案手法の再現率と適合率の両方が重要となり，再現率が高ければ高いほど，多くのノードとリンクを構築できたことになり，適合率が高ければ高いほど，意味ネットワークは Web コーパスに記述された行動と行動間の関係を正確に表していることになる．

提案手法では，行動と行動間の遷移の抽出精度（適合率）はそれぞれ 88.9% と 87.5%です．上記の twitter から抽出した文の評価結果によると，再現率が適合率より約 6.7 %下がってしまうため，行動と行動間の遷移の抽出再現率はそれぞれ 82.2%（88.9-6.7），80.8%（87.5-6.7）となる．つまり，提案手法では Web コーパスに記述された行動と行動間の遷移をそれぞれ 82.2%，80.8%で抽出でき，一定のノードとリンクの構築は可能だと考えられる．また図 21 に示すように，提案手法では，ノードとリンクをそれぞれ 88.9%（行動抽出の適合率）と 87.5%（行動間の遷移の適合率）で正確に表している．今後，より多くの行動と行動間の関係を意味ネットワークに加え，行動パターンを推薦できる可能性を上げていきたいと考えている．

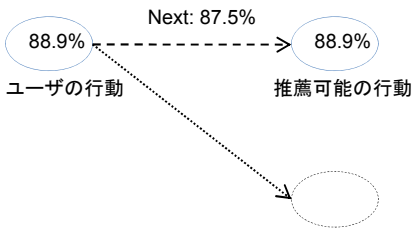

図 21 行動の推薦

5. 関連研究

本章では，まずマイクロブログに関する研究について検討する．次に，他の関連研究との比較を示す．

5·1 マイクロブログに関する研究

我々の調べた結果では，マイクロブログを対象にして人間行動の抽出・解析を行う研究は少なく，[Mor 10] と

[Nilanjan 09] だけがあった．

[Mor 10] は，twitter ユーザがどんなこと（自分の状況であるか，意見であるかなど）について発言するかを調査した論文である．調査方法は，“ Me now ”や“ Opinions/Complaints ”などの 9 つのカテゴリに基づいて，ユーザが発言したメッセージの内容を分類するというものである．調査結果よると，自分の現状について発言しているメッセージは全体の約 41%を占める．

[Nilanjan 09] は，人間の行動に着目して，twitter ユーザの興味をリアルタイムで把握する手法を提案している．この手法では，まずユーザの興味に関する「カテゴリワード」（例えば，movie, cinema, music, sports など）,「動詞」（例えば，watch, watching, party など）,「時間ワード」（例えば，tonight, today, weekend など）を作っておく．次に，これらのワードに基づいて，twitter メッセージに対して，（カテゴリワード，動詞）と（カテゴリワード，時間ワード）の共起頻度を計算する．共起頻度が高いものはユーザの興味があるものと判断する．例えば，“ movie ”（カテゴリワード）と“ go ”（動詞）と“ tomorrow ”（時間ワード）の共起頻度が高ければ，ユーザの興味は“ going to a movie tomorrow ”（明日，映画を見に行く）であると判断する．

この手法の利点は，twitter 文書の文法に依存せず，高速に結果を出力できるところにある．しかし，出力された結果の妥当性を評価していない．また，この手法は以下のような問題点があると考えられる．

(1) この手法では頻度が低い興味（行動）は，獲得できない．しかし，頻度が低い興味（行動）でも，重要な経験や意外性を含んでいる可能性は高いと言える．また，頻度が高い興味しか抽出できないので，この手法の再現率は低いと考えられる．

(2) “ today ”や“ tomorrow ”などのワードの頻度だけでは，興味（行動）をリアルタイムで把握できるとは言えない．これは，ユーザが発言したメッセージの時刻を考慮しなければ，いつに対しての“ today ”や“ tomorrow ”であるかが分からないためである．更に，メッセージに記述された行動主を把握していないので，誰の興味であるかも分からない．

(3) 頻度に基づくため，多くのデータ（メッセージ）から計算しないと正確な結果を得ることができない．

上記のように，我々の調べた限りでは，マイクロブログを対象として，人間の行動を抽出・解析する研究はまだ数が少なく，多くの問題点がある．しかし，タイムスタンプを利用した研究はいずれかにあると思われる．しかし，タイムスタンプに基づいて，行動間の関係を把握するには以下のような問題がある．

(1) 行動の発生がメッセージの発言された時刻と逆順である場合，行動間の遷移関係が誤って判断されてしまう．例えば， twitter から表 3 のような 2 つのメッセージを取得できるとする．メッセージ 1 では，（動作 = 行く，対象 = 秋葉原，時刻 = 来週）（行動 A) という行動を表す．メッセージ 2 では, (動作 = 食べる，対象 = ラーメン，時刻 = 昨日，場所 = 渋谷）（行動 B）という行動を表す．

Twitter のタイムスタンプ（18::00, 18:30）に基づくと，行動 A が起こった後に，行動 B が起こるという間違った判断になってしまう．しかし，我々の提案手法では，文中に表す時刻（来週と昨日）を解析すると，行動 B が起こった後に行動 A が起こるという正しい判断が可能である．つまり，行動間の遷移関係を正確に把握するためには，メッセージの内容を解析する必要があると思われる．

表 3 Twitter のメッセージ

No	Twitter に発言した時刻	行動を表す文
1	18:00	来週、秋葉原へ行く予定。
2	18:30	昨日，渋谷でラーメンを食べたよ。

(2) また，タイムスタンプを利用することで，ある程度，行動間の遷移（流れ）を推論できるが，行動間の因果関係の推論は非常に難しい．我々の提案手法では, “ ので ”, “ ため ”のような因果関係を表すパターンを訓練データの特徴モデルに加えることにより，行動間の因果関係を抽出できる．例えば, “ 俺は iPhone を買うために，秋葉原へ行ったよ ”の文に対して, “ 動詞ーために ”というパターンを利用することで，以下の出力を得ることができる．
→行動 A:（行動主 = 俺，動作 = 買う，対象 = iPhone）
→行動 B:（行動主 = 俺，動作 = 行った，対象 = 秋葉原）
行動 B BecauseOf 行動 A (行動 B が起こる原因は行動 A である).

更に，本論文では，あまり頻度の高くない行動（意外の行動，珍しい行動，貴重な経験等）に対しても，文に明確に記述されていれば抽出可能である．また, “ だから ”, “ したがって ”, “ そして ”, “ それから ”，などの接続詞を利用することで，文間に表す行動間の関係も把握できる．例えば, “ 俺は先週秋葉原へ行った。それから，アバターを見に行ったよ。”の文では, “ そして ”というパターンを用いると，以下の関係が分かる．
→行動 A:（行動主 = 俺，動作 = 行った，対象 = 秋葉原）
→行動 B：（行動主 = 俺，動作 = 見に行った，対象 = アバター）
行動 A Next 行動 B (行動 A が起こった後に，行動 B が起こった).

しかし，現時点では，暗黙的な行動やドキュメント間に表す行動の関係は抽出できていない．例えば, “ 俺は先週渋谷にある回転寿司店へ行った。すごく美味しかった

よ！”の文では，“ 寿司を食べた ”という行動が隠れており，抽出できない．今後，これらの課題を解決するために，推論や HTML 内の文書関係（リンク関係）の抽出，相関ルールの適用などを検討している．

5·2 他の関連研究

Web からの関係抽出と人間行動抽出に関しては，以下のような関連研究がある．以下に各研究の手法を説明し，本手法との比較を行う．

Web からの関係抽出の代表的な手法として，DIPRE[Brin 98]，SnowBall[Agichtein 00]，KnowItAll[Etzioni 04]，Pasca[Pasca 06], TextRunner[Banko 07]，O-CRF[Banko 08]（TextRunner の改善版）が挙げられる．

DIPRE，SnowBall，KnowItAll，Pasca はブートストラッピングを利用している．ブートストラッピングの利点は，単純なパターンマッチでは困難であった，情報を抽出するためのパターンを自動生成できることにある．一方で，ブートストラッピングの欠点としては，欲しい情報の周辺のパターンを誤って抽出すると，誤ったパターン周辺の欲しくない情報を抽出してしまうという問題がある．また，パターンを自動生成するための戦略がヒューリスティクスであり，そのため，意外性や発見性のあるパターンの生成が難しいことも欠点として挙げられる．これらの問題に加えて，次の 2 つの理由でブートストラッピングは文に現れる行動属性を抽出することが困難である．一つ目は文によって現れる行動属性の数と位置が変わるので，正規表現パターンの作成は非常に難しい．二つ目は行動のドメインを定義しておき，インスタンスを作成する必要がある．TextRunner は世界初の Open RE（Open Relation Extraction）*3のシステムであり，自己教師あり学習と Naive Bayes という分類手法を用いて，バイナリリレーションを抽出する．この手法では，O-CRF と同じようにエンティティを事前に判定しておき，リレーションがこれらのエンティティの間に現れる必要がある．また [Banko 08] の実験結果から，O-CRF と比べて TextRunner の抽出精度は低い（適合率　86.6%，再現率　23.2%）．

Web からの人間行動抽出の先行研究には，Perkowitz ら [Perkowitz 04]，川村ら [川村 08]，倉島ら [倉島 09] と Fukazawa ら [Fukazawa 09] の研究がある．

Perkowitz ら [Perkowitz 04] の手法は単純なキーワードマッチなので，作業の手順（料理の作り方など）を明示的に書いたウェブページにしか対応できない．また，行動属性間の係り受け関係が解析されていない．川村ら [川村 08] の手法では，行動オントロジーと対象トピックに関する情報（商品名など）のオントロジーを予め準備しておく必要があり，抽出精度はこれらのオントロジーに依存する．倉島ら [倉島 09] の手法では，ブログの日付情報から時刻を取得するので，行動文に表す時刻ではない可能性が高い．場所は，固有表現抽出器で“ 地名 ”,“ 組織 ”と判定される語なので，動作と係り受け関係がない可能性がある．対象と動作の抽出では，係り受けと各分析の自然言語処理ツール（JTAG[Fuchi 98]）を用いる．この方法は JTAG の精度に依存することとなる．また，助詞“ を ”と“ に ”が共にない文に対応できない．更に，Banko ら [Banko 08] が指摘するように，係り受け解析の自然言語処理ツールを直接用いてエンティティ（行動属性など）の相互関連を判定するのは Web コーパスに適切ではない．

Fukazawa ら [Fukazawa 09] の手法では，まず「ドメイン + 助詞（を，に) + 動詞」というパターンを用いて，検索エンジンでドメインと動詞を取得する．次に，Score(ドメイン，動詞) を計算し，10^{-5} より大きければこのドメインは対象，この動詞は動作として獲得する．

$$Score = \frac{Hits(\text{ドメイン } AND \text{ 動詞})}{Hits(\text{ドメイン})Hits(\text{動詞})} \quad (3)$$

この手法の利点は，検索エンジンだけ利用することで，対象と動作のペアを獲得できる．しかし，パターンを特定しており，再現率が非常に低いと考えられる．また，価値がある行動パターンでも，共起頻度が低ければ獲得できない．

以上の解析をまとめると，行動抽出の関連研究との主な比較は表 4 の通りである．

表 4 行動抽出の関連研究との比較

手法等	抽出可能属性	セットアップコスト	対応可能文	因果関係
Perkowitz ら	×	○	×	不可
Fukazawa ら	△	○	×	不可
Kurashima ら	○	○	△	N/A*4
Kawamura ら	△	△	○	不可
提案手法	○	○	□	可能

6. おわりに

本論文では，日本語 Web ページを対象とし，条件付確率場と自己教師あり学習を用いて，文に現れる人間行動の基本属性と行動間の遷移を自動的に抽出する手法を提案した．提案手法では，3·2 節 に示した日本語 Web ページにおける行動属性抽出の難しさと第 1 章に示した先行研究の問題点を解決でき，以下のような貢献がある．

- 人手でラベル編集，初期インスタンスの作成，行動のドメインの定義などの必要がなく，準備コストがかからない．
- 一回のテストでテストデータに現れる行動属性と行動間の遷移を漏れなく全て抽出でき，高い精度が得

*3 Open RE の概念は Banko ら [Banko 08] の研究グループにより定義されたものである．

*4 論文上では，行動間の因果関係の抽出を述べていない．自然言語処理ツールを用いて係り受けを上手く解析できると，ある程度行動間の因果関係を抽出できるが，解析時間がかかる．

られる（行動：88.9%，基本行動属性：90%以上，行動間の遷移：87.5%）．

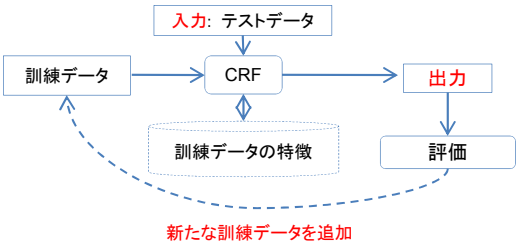

図 22　フィードバック: 出力を評価して新たな訓練データを追加

図 22 に示すように，我々は現在，複雑な文や正しい文法で記述されていない文に対して，抽出精度を向上するために，新たな訓練データの追加方法を検討している．今後の課題として，まず Web コーパスから大規模な行動を表す文を収集し，評価実験を行う．次に，HTML 内の文書関係，タイムスタンプ，相関ルールなどを考慮して，ドキュメント間に現れる行動間の関係の抽出手法を検討する．その後，人間の経験を対象として，Web コーパス全体からすべての行動属性と行動間の関係を抽出し，人間行動を表す意味ネットワークを構築する．そして，この意味ネットワークを参照して，ユーザの行動意図を把握し，賢い行動パターンを推薦する実世界指向エージェントの実現を目指す．

◇ 参　考　文　献 ◇

[Agichtein 00]　Agichtein, E. and Gravano, L.: Snowball: Extracting relations from large plain-text collections, in *Proc. ACM DL 2000* (2000)
[Banko 07]　Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O.: Open information extraction from the web, in *Proc. IJCAI2007*, pp. 2670–2676 (2007)
[Banko 08]　Banko, M. and Etzioni, O.: The Tradeoffs Between Traditional and Open Relation Extraction, in *Proc. ACL-08* (2008)
[Brin 98]　Brin, S.: Extracting Patterns and Relations from the World Wide Web, in *Proc. EDBT*, pp. 172–183 (1998)
[CoNLL]　CoNLL, : CoNLL 2000 shared task: Chunking
[Etzioni 04]　Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., S.Weld, D., and Yates, A.: Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison, in *Proc. AAAI-04* (2004)
[Forney 73]　Forney, G.: The viterbi algorithm, in *Proc. IEEE*, pp. 268–278 (1973)
[Fuchi 98]　Fuchi, T. and Takagi, S.: Japanese morphological analyzer using word co-occurence-JTAG, in *Proc. ACL-98*, pp. 409–413 (1998)
[Fukazawa 09]　Fukazawa, Y. and Ota, J.: Learning User's Real World Activity Model from the Web (2009)
[Google]　Google, : Google Maps API Services
[KDDI]　KDDI, : ユビキタスネットワーク技術の研究開発 ～ ケータイ de ライフログ ～
[Kudo 02]　Kudo, T. and Matsumoto, Y.: Japanese Dependency Analysis using Cascaded Chunking, in *Proc. CoNLL 2002*, pp. 63–69 (2002)
[Kudo 04]　Kudo, T., Yamamoto, K., and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphologiaical Analysis, in *Proc. EMNLP2004*, pp. 230–237 (2004)
[Lafferty 01]　Lafferty, J., McCallum, A., and Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in *Proc. ICML2001* (2001)
[Matsuo 07]　Matsuo, Y., Okazaki, N., Izumi, K., Nakamura, Y., Nishimura, T., and Hasida, K.: Inferring long-term user properties based on users' location history, in *Proc. IJCAI2007*, pp. 2159–2165 (2007)
[McCallum 03]　McCallum, A. and Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons (2003)
[Mor 10]　Mor, N., Jeffrey, B., and Chih-Hui, L.: Is it Really About Me. Message Content in Social Awareness Streams, in *Proc. CSC 2010* (2010)
[NikkeiBP]　NikkeiBP, : 日本企業よ，Google の先を走れ！
[Nilanjan 09]　Nilanjan, B., Dipanjan, C., Koustuv, D., Anupam, J., Sumit, M., Seema, N., Angshu, R., and Sameer, M.: User Interests in Social Media Sites: An Exploration with Micro-blogs, in *Proc. CIKM 2009* (2009)
[NTTDocomo]　NTTDocomo, : 情報大航海プロジェクト マイ・ライフ・アシストサービス概要
[Ozok 09]　Ozok, A. A. and Zaphiris, P.: *Online Communities and Social Computing*, Third International Conference, OCSC 2009, Held as Part of HCI International 2009, San Diego, CA, USA. Springer, ISBN-10: 3642027733 (2009)
[Pasca 06]　Pasca, M., Lin, D., Bigham, J., Lifchits, A., and Jain, A.: Organizing and Searching the World Wide Web of Facts - Step One: the One-Million Fact Extraction Challenge, in *Proc. AAAI-06*, pp. 1400–1405 (2006)
[Peppers 99]　Peppers, D. and Rogers, M.: *The One to One Fieldbook*, Broadway Business ISBN-10: 038549369X (1999)
[Perkowitz 04]　Perkowitz, M., Philipose, M., and J.Patterson, K. F. D.: Mining Models of Human Activities from the Web, in *Proc. WWW2004* (2004)
[Poslad 09]　Poslad, S.: *Ubiquitous Computing Smart Devices, Environments and Interactions*, Wiley, ISBN: 978-0-470-03560-3 (2009)
[Sha 03]　Sha, F. and Pereira, F.: Shallow parsing with conditional random fields, in *Proc. HLTNAACL*, pp. 213–220 (2003)
[Twitter]　Twitter, I.: http://twitter.com
[Wikipediaa]　Wikipedia, : Wikipedia, Category:人物
[Wikipediab]　Wikipedia, : Wikipedia, Category:人名
[川村 08]　川村隆浩, 山崎智弘, 長野伸一, 溝口祐美子, 飯田貴之：CGM からのユーザ行動マイニングの提案, in *The 22nd Annual Conference of the Japanese Society for Artificial Intelligence* (2008)
[倉島 09]　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, 電子情報通信学会論文誌, 情報・システム, pp. 301–310 (2009)

〔担当委員：服部 宏充〕

2009 年 12 月 25 日　受理

◇ 付　　録 ◇

A.　付録 A　評価実験のデータ

- 実験データ, `http://docs.google.com/View?id=dftc9r33_901fjwmkwfk`
- 実験データのリソース 1, `http://www.geocities.jp/niwasaburoo`
- 実験データのリソース 2, `http://ja.wikipedia.org/wiki/アルベルト・アインシュタイン`
- 実験データのリソース 3, `http://ameblo.jp/aromafitness/entry-10299513069.html`

B.　付録 B　評価実験の結果

- 実験の結果, `http://docs.google.com/View?id=dftc9r33_902cxq5r7cr`

C.　付録 C　O-CRF の実験データ

- O-CRF の実験データ，http://www.cs.washington.edu/research/knowitall/hlt-naacl08-data.txt(＜ $p1$ ＞と ＜ /$p1$ ＞ の中に入る単語はエンティティ 1，＜ $p2$ ＞と ＜ /$p2$ ＞ の中に入る単語はエンティティ 2 である）.

D.　付録 D　前処理の評価実験

- 集合 A，http://docs.google.com/View?id=dftc9r33_1958hhmmhjhj
- 捨てられた文，http://docs.google.com/View?id=dftc9r33_1959ff2b7fxp
- 集合 C，http://docs.google.com/View?id=dftc9r33_1960c55rsncv

著　者　紹　介

グェン　ミン　テイ

2009 年電気通信大学大学院情報システム学研究科修士課程修了．現在，同大学院博士課程　社会知能情報学専攻在学．主にユーザの状況に応じた行動推薦に関する研究に従事．

川村　隆浩（正会員）

1992 年早稲田大学理工学部電気工学科卒業．1994 年同大大学院理工学研究科電気工学専攻 修士課程了．同年，(株) 東芝入社．現在，同社研究開発センター 主任研究員．工学博士．2001–2002 年米国カーネギ・メロン大学 ロボット工学研究所 客員研究員．2003 年より電気通信大学大学院情報システム学研究科 客員准教授．2007 年より大阪大学大学院 工学研究科 非常勤講師．主としてマルチエージェントシステム，セマンティック Web の研究・開発に従事．情報処理学会会員．

中川　博之

1974 年生．1997 年大阪大学基礎工学部情報工学科卒業．同年鹿島建設（株）に入社．2007 年東京大学大学院情報理工学系研究科修士課程修了，2008 年同大学院博士課程中退．同年より電気通信大学助教，現在に至る．エージェントおよび自己適応システム開発手法の研究に従事．情報処理学会，電子情報通信学会，IEEE CS 各会員．

田原　康之

1966 年生. 1991 年東京大学大学院理学系研究科数学専攻修士課程修了. 同年 (株) 東芝入社. 1993〜1996 年情報処理振興事業協会に出向. 1996〜1997 年英国 City 大学客員研究員. 1997〜1998 年英国 Imperial College 客員研究員. 2003 年国立情報学研究所入所. 2008 年より電気通信大学准教授. 博士 (情報科学)(早稲田大学). エージェント技術，およびソフトウェア工学などの研究に従事. 情報処理学会，日本ソフトウェア科学会会員.

大須賀　昭彦（正会員）

1981 年上智大学理工学部数学科卒．同年 (株) 東芝入社．同社 研究開発センター，ソフトウェア技術センターなどに所属．1985〜1989 年 (財) 新世代コンピュータ技術開発機構 (ICOT) 出向．2007 年より，電気通信大学大学院情報システム学研究科 教授．工学博士 (早稲田大学)．主としてソフトウェアのためのフォーマルメソッド，エージェント技術の研究に従事．1986 年度情報処理学会論文賞受賞．情報処理学会，電子情報通信学会，日本ソフトウェア科学会，IEEE CS 各会員．