1D1-3

CGMからの自己教師あり学習と条件付確率場を用いた人間行動マイニング

グェン ミン テイ Nguyen Minh The
川村 隆浩 Kawamura Takahiro
中川 博之 Nakagawa Hiroyuki
田原 康之 Tahara Yasuyuki
大須賀 昭彦 Ohsuga Akihiko

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

The goal of this paper is to describe a method to automatically extract *all* basic attributes namely *actor*, *action*, *object*, *time* and *location* which belong to an activity, in each sentence retrieved from Japanese CGM (consumer generated media). Previous work had some limitations, such as high setup cost, inability of extracting all attributes, limitation on the types of sentences that can be handled, and insufficient consideration of interdependency among attributes. To resolve these problems, this paper proposes a novel approach that treats the activity extraction as a sequence labeling problem, and automatically makes its own training data. This approach has advantages such as *domain-independence*, *scalability*, and *unnecessary hand-tagged data*. Since it is unnecessary to fix the positions and the number of the attributes in activity sentences, this approach can extract *all* attributes by making *only a single pass* over its corpus. In an experiment, this approach achieves high precision (activity: 88.87%, attributes: over 90%).

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 [7] とソーシャルコンピューティング [8] 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 [12]）や他人の経験に基づき最適なアドバイスをする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．

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，センサーイベントやWeb から行動データを収集し，行動属性の抽出を行う先行研究 [21, 22, 6, 19, 1] がある．しかし，イベントデータにノイズが多いことや [21, 22]，抽出のための準備コストが大きいこと [6]，抽出できる行動属性が少ないこと [1, 6]，適用可能な文の種類が少ないこと [1, 19]，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 [1, 19] などといった問題がある．

そこで本論文では，日本語の CGM（ブログ，Twitter 等）から取得した「行動を表す文」を対象とし，文中に現れる基本行動属性（行動主，動作，対象，場所，時間）の自動抽出手法を提案する．提案手法はバイナリリレーション抽出の最先端技術である O-CRF[4] の考え方をベースとし，条件付確率場と自己教師あり学習*1を利用する．まず，少量のサンプルデータから各文に現れる行動属性を抽出し，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）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（CGM から取得した行動を表す文）の行動属性を抽出する．

本論文は，次のような構成をとる．第 2. 章では，条件付確率場と自己教師あり学習を用いた行動属性の自動抽出手法を提案する．第 3. 章では，評価実験と考察を述べる．第 4. 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 5. 章では，今後の課題と合わせて本論文をまとめる．

連絡先: {minh,kawamura,nakagawa,tahara,akihiko}@ohsuga.is.uec.ac.jp

*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，自己教師あり学習と呼ぶ．

2. 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

2.1 条件付確率場とは

条件付確率場（Conditional Random Fields）とは，John D. Lafferty ら [13] が提案した系列ラベリング問題に適用した識別モデル（discriminative model）である．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 は柔軟な素性設計を可能にし，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 [13]，テキストチャンキング [14]，固有表現抽出 [15]，形態素解析 [16] 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．

2.2 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 ”行動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，これらの属性にそれぞれ Who, Action，What，Where，When というラベルを付ける．

「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．

本論文の課題は，日本語の CGM から取得した「行動を表す文」に現れる行動の基本属性を自動的に抽出することである．例えば,“ これから，秋葉原へ行くよ ”の文に現れる基本属性は「動作」(“ 行く ”) と「対象」(“ 秋葉原 ”) である．

2.3 行動属性抽出の難しさ

日本語 CGM から取得した文に現れる行動属性の抽出については，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，日本語の文において，CRF を直接適用することは困難である．

2. CGM から取得する文では，複雑かつ文法的に正しくない文が多い．そして，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も多い．

3. O-CRF[4] は英語 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 1 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．そして，エンティティを事前に判定しておくので，抽出対象はリレーションだけである．一方，提案手法は，文に現れるすべての行動属性を抽出する必要がある．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 の解決課題よりも本論文が解決する課題が困難であると考えられる．

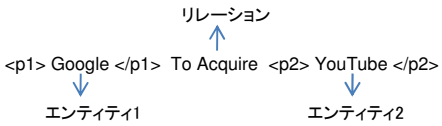

図 1: O-CRF の実験データ

2.4 提案手法のアーキテクチャ

CGM 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 CGM から取得した文中に現れる行動属性を抽出する．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 2 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 2 の I）と行動属性の自動抽出モジュール（図 2 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い．まず，Wikipedia の人物カテゴリ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を抽出し．最後に，抽出した結果を合わせて，練データを自動的に作成する．

行動属性の自動抽出モジュールは，CGM 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を自動的に抽出する．

各モジュールの詳細を以下に示す．

2.4.1 訓練データの自動生成モジュール

以下,“ 太郎は歯を磨く ”という例文を用い，訓練データの自動作成方法を説明する．

1. Mecab[16] で形態素解析を行い，文の単語と単語の品詞記号を取得する（図 2 の I.1）．

2. Cabocha[18] 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する（図 2 の I.2）．

3. 形態素解析結果に加えて,“ 今 ”,“ 曜日 ”などのような文字列を含む時間表現を時刻として判定する．そして，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時間表現（時刻，場面）を抽出する（図 2 の I.3）．

4. 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 2 の I.4）．抽出精度を向上するために，形態素解析結果に加えて，Google Map API を用いて場所を判定する．

5. 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 の人名カテゴリを活用して文の人名を検出する（図 2 の I.5）．

6. 日本語の構文リスト (NP ヲ VP，NP ガ NP 二 VP など) を用い，これらの NP，VP はどれが行動主，動作，対象であるかを判定する（図 2 の I.6）．

7. 以上の解析結果を合わせて訓練データを自動的に作成する（図 2 の I.7）．例文の訓練データは図 3 のようになる．

太郎	44	B-Who
は	16	O
歯	38	B-What
を	w	O
磨く	v	B-Action

図 3: 例文の訓練データ

2.4.2 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. 図 4 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 2 の II.1）．

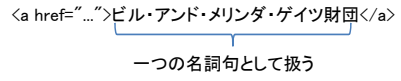

図 4: HTML タグを解析して名詞句を取得

2. 文の複雑な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 5 のようになる（図 2 の II.2 と II.3）．

3. 条件付確率場とテンプレートファイル（図 2 の T）を用いて，訓練データの特徴モデルに基づき，テストデータの行動属性を自動的に抽出する（図 2 の II.4）．

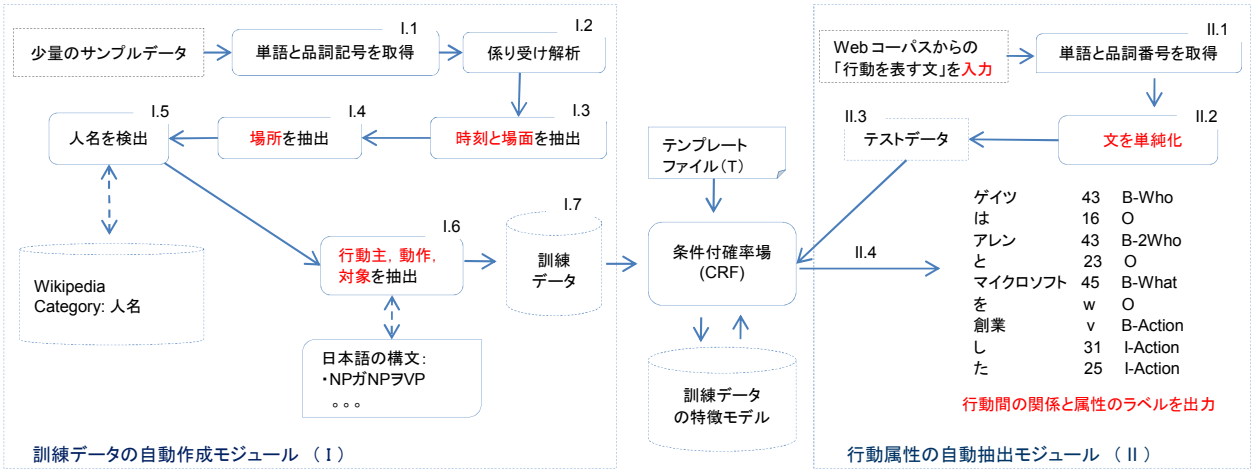

図 2: 提案手法のアーキテクチャ

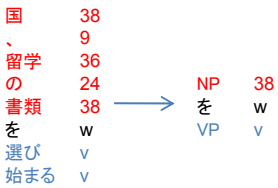

図 5: 複雑な名詞句と動詞句を単純化

2.4.3 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 6 のようになる．

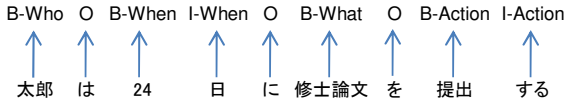

図 6: 行動属性を系列ラベリングで表現

2.1 節に述べたように系列ラベリング問題を適用する学習モデルのうち条件付確率場が高い精度を得るので，本手法はこれを利用する．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 のウィンドウを採用する．

3. 評価実験

3.1 実験

提案手法を用いることで，先行研究の問題点と 2.3 節にに述べた課題を解決できたかどうかに加えて，行動の基本属性の抽出精度を明らかにするために，我々は評価実験を行った．評価実験のデータセットは「行動を表す」533 文[*2]である．これは CGM 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．

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を正しく抽出することができた場合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（533）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[*3]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	710	631	88.87
行動主	196	182	92.86
動作	710	693	97.61
対象	509	479	94.11
時間	173	165	95.38
場所	130	120	92.31

3.2 考察

実験結果により， 本手法では，日本語の CGM 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）を自動的に抽出でき，高い抽出精度を得た．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，533 文の抽出時間は約 0.27 秒であったが，一方で，Cabocha を用いて，この 533 文の係り受けのみを解析した場合の処理時間は約 46.45 秒であり，提案手法の 172 倍であった．従って，提案手法は CGM 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，属性の数と位置を固定する必要がないため，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 CGM 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．

*2 http://docs.google.com/View?id=dftc9r33_1077g63vrjc5

*3 http://docs.google.com/View?id=dftc9r33_1078cr9hd3mt

- 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 CGM 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，2.3 に示した日本語の CGM 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. 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．
3. 日本語の構文パターン（ヒューリスティクス）を用いることで，属性の数と位置を固定する必要がない．

4. 関連研究

Web コーパスから人間行動抽出の先行研究には，Perkowitz ら [1]，川村ら [6] と倉島ら [19] の研究がある．

Perkowitz ら [1] 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 [6] 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 [19] 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[17]）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 [5] 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 コーパスに適用ではない．

5. おわりに

本論文では，日本語の CGM 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を自動的に抽出する手法を提案した．提案手法では，2.3 節に示した日本語 CGM 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．
- 一回の実行でテストデータに現れる行動属性を漏れなく全て抽出でき，高い精度が得られる（行動：88.87%，基本行動属性：90%以上）．
- アーキテクチャのスケーラビリティが高く，膨大かつ多様性を持つ CGM コーパスに対応できる．

今後の課題として，まず CGM 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，文内だけでなく，HTML 内の文書関係を考慮して，文間と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，CGM 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，各消費者の行動意図を把握し，最適な行動パターンを推薦する．

参考文献

[1] Perkowitz, M., Philipose, M., Fishkin, K., Patterson, D. J.: Mining Models of Human Activities from the Web. In Proc. WWW2004 (2004)

[2] Pasca, M., Lin, D., Bigham, J., Lifchits, A., 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)

[3] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D., Yates, A.: Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison. In Proc. AAAI-04 (2004)

[4] Banko, M., Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction. In Proc. ACL-08 (2008)

[5] Michele Banko: Open Information Extraction for the Web. PhD thesis, University of Washington (2009)

[6] Kawamura, T., The, N. M., and Ohsuga, A.: Building of human activity correlation map from weblogs. In Proc. ICSOFT (2009)

[7] Poslad, S.: Ubiquitous Computing Smart Devices, Environments and Interactions. Wiley, ISBN: 978-0-470-03560-3 (2009)

[8] Ozok, A. A., 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)

[9] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the Web. In Proc. IJCAI2007, pp. 2670-2676 (2007)

[10] Brin, S.: Extracting Patterns and Relations from the World Wide Web. In WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT '98, Valencia, Spain, pp.172-183 (1998)

[11] Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections. In Proc. ACM DL 2000 (2000)

[12] Peppers, D., Rogers, M.: The One to One Future. Broadway Business, ISBN-10: 0385485662 (1996)

[13] Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields Probabilistic models for segmenting and labeling sequence data. In Proc. ICML, pp.282-289 (2001)

[14] Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In Proc. HLTNAACL, pp.213-220 (2003)

[15] McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons. In Proc. CoNLL 2003 (2003)

[16] Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphologiaical Analysis. IPSJ SIG Notes pp.89-96 (2004)

[17] Fuchi, T., Takagi, S.: Japanese morphological analyzer using word co-occurence-JTAG, In Proc. ACL-98, pp. 409-413 (1998)

[18] Kudo, T., Matsumoto, Y.: Japanese Dependency Analysis using Cascaded Chunking. In Proc. CoNLL 2002, pp.63-69 (2002)

[19] Kurashima, T., Fujimura, K., Okuda, H.: Discovering Association Rules on Experiences from Large-Scale Weblogs Entries. ECIR 2009, pp.546-553 (2009)

[20] Forney, G.D.: The viterbi algorithm. Proceedings of the IEEE, pp.268–278 (1973)

[21] NTT Docomo, Inc.: 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 (2008).

[22] KDDI, Corp.: 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 〜 (2008).