1D2-3

機械学習とユーザ知識を用いたイベント情報の構造化

Extraction of Structured Information by Machine Learning using Community Information

森近 憲行[*1] Noriyuki Morichika　濱崎 雅弘[*2] Masahiro Hamasaki　亀田 尭宙[*3] Akihiro Kameda　大向 一輝[*4] Ikki Ohmukai　武田 英明[*4] Hideaki Takeda

[*1]東京大学大学院工学系研究科 School of Engineering, The University of Tokyo

[*2]産業技術総合研究所 National Institute of Advanced Industrial Science and Technology

[*3]東京大学大学院情報理工学系研究科 Graduate School of Information Science and Technology, The University of Tokyo

[*4]国立情報学研究所 National Institute of Informatics

In this paper, we describe our approach for information extraction from documents, which is based on supervised machine learning and collective intelligence. This approach is aimed at redeeming each method, because each method has merits and demerits. It provides various ways for users to input data to improve information extraction. Users can add not only supervised data but also a rule to extract values for a set of attributes. Various ways to input data allows many users to add a lot of data for improvement and machine learning can reduce noise of data input by users. We implemented it in event-information extraction system, and the experimental result shows effectiveness in correctness and convenience.

1. はじめに

ウェブの普及に伴って，ウェブ上には膨大なドキュメントが蓄積されてきている．これらのドキュメントは，HTML というマークアップ言語の文法に沿って記述されるが，HTML によるマークアップは段落の構成やフォントの指定など，ドキュメントを人間が理解しやすいようにするものが多い．このように，現在のウェブは人が読むためのドキュメントが中心で構成されていることから，プログラムなどでウェブ上のドキュメントを，記述されている内容を考慮して大量に処理することができないという問題がある．これらのドキュメントを機械可読なものとするためには，ドキュメントから情報抽出を行い，構造化されたドキュメントに変換する必要性がある．

従来から研究がおこなわれている情報抽出の手法としては，人手で抽出規則を作成し抽出を行う方法や機械学習などの方法で抽出規則を自動作成し抽出を行う方法がある [1] [2]．これらの方法にはそれぞれ高い抽出精度が実現可能，一度分類器を構築してしまえばそれ以降保守の必要性はあまりない，という利点がある．しかしこれらの方法には，前者には膨大な抽出規則を作成する必要がある，後者には良い精度で抽出を行うためには事前にたくさんの教師データを用意する必要がある，といった欠点がそれぞれあった．

そこで我々は，機械学習による抽出とシステム利用者の集合知を組み合わせた，多段階の情報抽出手法を提案している．我々の提案する情報抽出手法では，大小さまざまな大きさの負荷のユーザ参加の方法を用意することでより多くの知識を集めることができることや，ユーザから得られる知識の再利用性を高めることができるなどの利点がある．音声認識分野において，後藤らは音声情報検索システム「PodCastle」において音声認識の結果をユーザに開示し，認識誤りを訂正する協力をしてもらうことで音声認識技術を発展させていくことを提案している [5]．

本稿では，我々の提案手法を実装したイベント情報構造化システムを用いて，提案手法の評価について検討する．被験者実験では，用意したさまざまなバリエーションのユーザ参加の負荷の大きさを計測し，実験で得られたユーザ知識をシステムに反映した結果，情報抽出の精度が上昇したことを示す．

以下，2. 章で提案手法について，3. 章でイベント情報を対象に実装したシステムについて説明し，4. 章で被験者実験の目的と実験方法，5. 章で実験の結果と考察について議論する．最後に，6. 章で本稿のまとめと今後の課題について述べる．

連絡先: 森近 憲行，東京大学大学院工学系研究科，東京都文京区本郷 7-3-1

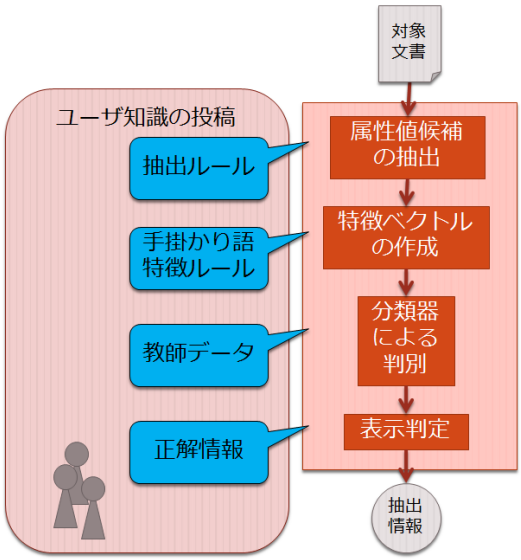

図 1: 情報抽出の流れ．

2. 提案手法

我々が提案する手法では，情報抽出を複数段階に分けて行い，情報抽出の各段階にユーザが投稿する情報や知識を反映させる．情報抽出の全体の流れを図 1 に示す．情報抽出を段階

図 2: イベント詳細画面.

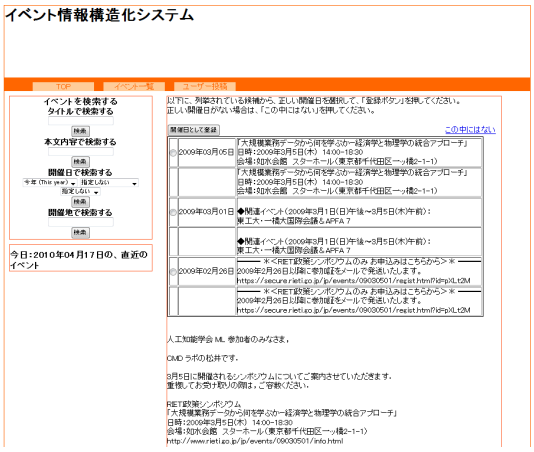

図 3: ユーザ投稿画面.

的に行うことで，さまざまなユーザ参加の方法を用意することが可能となるほか，ユーザから得た知識の再利用性を高めることが可能となる．この複数段階による情報抽出は，渡部らが提案する手法を参考にしている [4]．また，機械学習にユーザ知識を反映することで，ユーザ知識の信頼性を担保することができ，効果的にユーザ知識をシステム全体の精度向上に貢献させることができる．以下図 1 の情報抽出の流れについて説明する．

まず，情報抽出を対象とするドキュメントから抽出したい情報の候補（以下、属性値候補と呼ぶ）を抽出する．次に，得られた属性値候補をそれぞれ特徴ベクトルに変換する．そして，機械学習によって構築した分類器によって，これらの特徴ベクトルがそれぞれ属性値，すなわち抽出したい情報であるかどうかを分類する．最後に，分類器によって得られた分類結果に対して，ユーザからの投稿情報を反映して最終的な表示結果を判定する．

この情報抽出の流れの各段階に対して，システムの利用者はさまざまな方法で投稿を行うことができる．例えば属性値候補の抽出の段階には属性値候補の抽出ルールを，特徴ベクトルの作成の段階には特徴ベクトルを作成するための抽出ルールやキーワードなどの手掛かり語を，分類器による判別の段階には分類器を構築するための教師データを，最終的な表示判定を行う段階では分類器による誤分類を正すための正解情報を，それぞれ投稿することができる．

3. 実装システム

本章では，提案手法を実装した「イベント情報構造化システム」について説明する．「イベント情報構造化システム」は人工知能学会のメーリングリスト*1 で告知されるイベント情報を対象に，提案手法を実装した情報構造化システムであり，抽出を行ったのはイベントのタイトル情報と開催地情報と開催日情報である．図 2 に実装したシステムであるイベントについての詳細表示画面のスクリーンショットを，図 3 に開催日情報の修正情報投稿画面のスクリーンショットを示す．このシステムでは，イベントの開催日で検索を行ったり，近日中に行われるイベント一覧を表示することができる．システムを使っているユーザは，イベントを探す途中で抽出情報の間違いに気づいた場合にはユーザ参加フォームから修正情報を投稿したり，

*1 http://www.ai-gakkai.or.jp/jsai/ml/

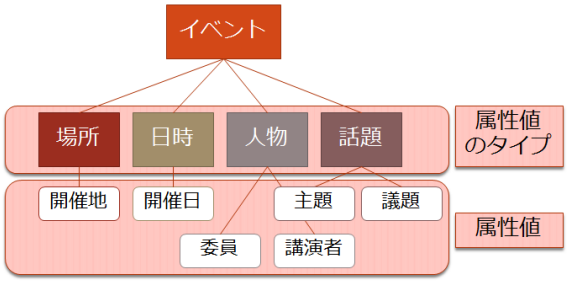

図 4: イベント情報の構造.

抽出ルールを投稿したりすることができる．タイトル情報については，今回扱ったメールドキュメントの特性から，メールのタイトルをイベントのタイトルとした．以下，提案手法をどのように実装したかについて説明する．

3.1 対象としたドキュメントおよび問題設定

本稿では，メールドキュメントからイベント情報を抽出するという問題を扱った．対象としたメールドキュメントは，人工知能学会のメーリングリストであり，このメーリングリストで告知される学会の情報やシンポジウムの情報の構造化を行った．このメーリングリストを対象とした理由は，メーリングリストの管理者によってメールの内容が原稿募集なのかイベントの参加募集なのか，イベントの参加募集のメールならばイベント開催日はいつなのか，といったタグが事前に付与されており，情報抽出システムによる情報抽出の精度計算が容易に行えることが挙げられる．問題設定は，このメーリングリスト上のメールのうち原稿募集およびイベントの参加募集のものを対象に，メールドキュメントを改行文字列で分割し，各行に抽出したいイベントに関する情報を含んでいるか否かを判定する問題とした．

3.2 属性値候補の抽出

図 4 に，抽出を行いたいイベント情報の構造の例を示す．例えばイベント情報には場所情報や日時情報，人物情報，話題情報などが属性値のタイプとしてあり，それぞれの下位情報として，より具体的な，最終的に抽出を行いたい情報が属性値として考えられる．例えば日時に関する情報として開催日情報など

があり，人物に関する情報としては委員情報や講演者情報などの複数の属性項目がある場合もある．この段階では，ドキュメントの中から抽出したいイベント情報の属性値のタイプ，に当たる情報を属性値候補として抽出する．抽出ルールは，正規表現にマッチするか否か，特定の単語を含むか否かなど，それぞれ抽出を行いたい属性値候補に合った方法をとる．今回の実装では，日付情報の抽出には用意した，もしくは投稿された正規表現を各行に適用し，正規表現がマッチした行を属性値候補とした．地名情報の抽出は実装せず，すべての行を属性値候補とした．ユーザが投稿した抽出ルールも属性値候補の抽出に利用する．

3.3 特徴ベクトルの作成

ドキュメントから抽出されたそれぞれの属性値候補を特徴ベクトルに変換する．その変換ルールは以下の通りである．なお，形態素解析にはフリーの形態素解析エンジン「MeCab」[*2]を用いた．特徴ベクトル次元の作成は，山田らの手法を参考にしている [3].

1. 属性値候補の行と，本来のドキュメント中においてその前にある行の末尾 3 文字とその後にある行の先頭 3 文字を結合し，文字列 A を得る．A をもとに，特徴ベクトル $\mathbf{a}$ を作成する．

2. A から得られる特徴ベクトル $\mathbf{a}$ は，3 種類の特徴ベクトル $\mathbf{w}$，$\mathbf{k}$，$\mathbf{r}$ から成り，$\mathbf{a} = \{\mathbf{w}, \mathbf{k}, \mathbf{r}\}$ であらわされる．$\mathbf{w} = (w_1, w_2, ..., w_l)$ は単語ベクトルで，A が単語 W_i を文字列内に含めば $w_i = 1$，含まなければ $w_i = 0$ となる次元を持つベクトルである．単語 W_i はドキュメント集合全体において 2 回以上出現する名詞のみに限定した．$\mathbf{k} = (k_1, k_2, ..., k_m)$ はキーワードベクトルで，A が単語 K_i を文字列内に含めば $k_i = 1$，含まなければ $k_i = 0$ となる次元を持つベクトルである．単語 K_i は事前に用意したものおよびユーザが投稿したものを用いた．$\mathbf{r} = (r_1, r_2, ..., r_n)$ は抽出ルールベクトルで，A に抽出ルール R_i を適用し，マッチするもしくは真の値を返すならば $r_i = 1$，マッチしないもしくは偽の値を返すならば $r_i = 0$ となる次元を持つベクトルである．抽出ルール R_i は開催日であれば事前に用意した正規表現とユーザが投稿した正規表現，開催地であれば事前に用意した抽出ルールとユーザが投稿した抽出ルールを用いた．

3.4 分類器による判別

属性値候補から得られた特徴ベクトルを，機械学習アルゴリズムで作成した分類器によってそれぞれが属性値であるか否かを分類する．機械学習アルゴリズムはサポートベクトルマシンを用いた．分類器の作成には，事前に用意した教師データの他に，ユーザが投稿した正誤情報も教師データとして利用する．

3.5 表示判定

分類器での分類は誤って分類が行われる可能性も高い．ここまで全体的にユーザからの投稿の信頼性を担保するために分類器の教師データとしてユーザ投稿を利用していたが，ユーザから多くの投稿が行われている行に対しては，ユーザ投稿によって得られた知識の信頼性は高いと判断し，分類器での分類結果とユーザ知識を比較して最終的な分類判定を行う．

*2 http://mecab.sourceforge.net/

表 1: 開催地情報抽出の実験前後での抽出精度.

	実験前	実験後
再現率（%）	60.77	64.09
適合率（%）	72.37	80.00
F 値	66.07	71.17

表 2: 開催日情報抽出の実験前後での抽出精度.

	実験前	実験後
再現率（%）	17.94	20.38
適合率（%）	17.67	19.97
F 値	17.80	20.17

4. 被験者実験

本稿の被験者実験の目的は，ユーザから得た情報を機械学習を通してシステムに反映することで情報抽出の精度がどのように変化するのか，さまざまなユーザ参加を用意することでユーザはどのようなふるまいをし，ユーザへの負荷はどのようになっているのか，を検証することである．

実験は、3 種類の実験とアンケートを 8 名の被験者に遠隔でウェブブラウザ上で行ってもらった．実験 1 ではキーワード投稿を除く 6 種類のユーザ参加についてそれぞれウェブブラウザ上で説明し，5 回ずつ修正および投稿を行う問題を解いてもらった．実験 2 では自由にシステムを操作してもらい 5 回のユーザ参加を自由に行ってもらった．実験 3 ではイベント検索という実際にこのシステムを利用するような状況に近いタスクを与え，タスクをこなす途中で間違いを発見した場合にユーザ参加を行ってもらった．

5. 実験の結果と考察

5.1 得られたデータによる精度向上

実験で得られたデータの結果と考察を行う．まず、ユーザ知識による精度の向上についてだが，実験で得られたデータを全て教師データ，特徴ベクトル作成ルールとして追加して学習器を再構築し，追加の前後で情報抽出の精度を計算したものを表 1 および表 2 に示す．追加したデータは，開催地情報の教師データ 31 件，開催日情報の教師データ 31 件，開催地情報の特徴ベクトル作成ルール 0 件，開催日情報の特徴ベクトル作成ルール 3 件である．評価には，再現率，適合率および F 値を用いた．表より，開催地，開催日どちらの抽出においても，再現率，適合率，F 値全ての値がデータ追加前よりもデータ追加後のほうが高いことが分かる．このことより，開催地，開催日いずれも，ユーザ知識を追加することで情報抽出精度は向上したといえる．開催日の抽出精度が開催地の抽出精度に比べて低いことについてだが，日付情報には開催日のほかにも参加申込締切日や投稿締め切り日など似たような記述が複数存在することが影響していると考えられる．

次に，開催日情報の抽出において教師データ，特徴ベクトル作成ルールをそれぞれ別々に追加した場合の情報抽出精度の変化の結果を表 3 に示す．この表より, 教師データのみを追加した場合，特徴ベクトル作成ルールのみを追加した場合でも，追加をしなかった場合に比べて情報抽出精度は上がっているこ

表 3: 開催日情報抽出における，追加データの種類と精度向上の関係.

	追加なし（実験前）	教師データのみ追加（31 件）	ルールのみ追加（3 件）	両方追加（31+3 件）
再現率（%）	17.94	19.16	18.46	20.38
適合率（%）	17.67	18.77	18.18	19.97
F 値	17.80	18.96	18.32	20.17

表 4: アンケート結果：負荷の大きさ.

ユーザ参加の種類	実験 1．での平均所要時間（秒）	「大きな負担だった」「やや大きな負担だった」の回答の割合（%）
開催地：候補の中からの選択	11.3	28.6
開催地：文書全体からの選択	14.3	42.8
開催地：プログラムコードの投稿	189.8	85.7
開催日：候補の中からの選択	24.9	28.6
開催日：文書全体からの選択	31.5	42.8
開催地：正規表現の投稿	132	85.7

とが分かる．しかし，両方追加した場合の抽出精度が一番高くなっていることから，教師データの投稿，特徴ベクトル作成ルールの投稿どちらも抽出精度の向上に貢献するが，両方追加した場合が一番抽出精度の向上が大きく，どちらのデータも効率的にユーザから収集する必要性が示されたといえる．また，追加データ 1 件当たりの F 値上昇量は教師データが 0.07，特徴ベクトル作成ルールが 0.17 であり，データ 1 件当たりの情報抽出精度への貢献は教師データよりも特徴ベクトル作成ルールのほうが大きいと考えられる．

5.2 ユーザへの負荷と動機

次に，ユーザへの負荷とユーザ参加の動機について，アンケートの結果を用いて考察する．開催日のキーワード投稿を除く 6 種類のユーザ参加方法について，問題画面を示してからユーザが投稿情報を選択して投稿を行うまでの平均所要時間，ユーザ参加の負担について「大きな負担だった」「やや大きな負担だった」「どちらとも言えない」「あまり大きな負担ではなかった」「大きな負担ではなかった」の 5 段階評価で質問したうち，負荷が大きいという「大きな負担だった」「やや大きな負担だった」の 2 つの回答の合計割合を表 4 に示す.

表 4 より，開催地，開催日どちらにおいても，投稿までの平均所要時間，負荷が大きいという回答の割合それぞれの値について，候補からの選択よりもメール全体からの選択の方が，メール全体からの選択よりもプログラムコードや正規表現などの抽出ルールの投稿の方が値が大きいことが分かる．これより開催日情報，開催地情報それぞれについて，ユーザ参加の負荷の大きさは，小さい順から候補からの選択，メール全体からの選択，プログラムコードや正規表現などの抽出ルールの投稿，となっている，と考えられる．また，表 4 よりルールの投稿の負荷は他のユーザ参加に比べて極めて大きく，あまり多くのユーザはルールの投稿によるユーザ参加は実行してくれないことが考えらる．しかし，5.1 節で示したように，抽出ルールの方が教師データよりもデータ 1 件当たりの抽出精度への貢献は大きいことから，抽出ルールの収集を効率的にするために，抽出ルール投稿のユーザ参加の負荷軽減を行う必要がある．

6. おわりに

本稿では，機械学習と集合知を組み合わせた情報抽出手法を提案し，学会イベント情報を対象にイベント情報の構造化システムを実装した．また被験者実験を行い，複数のユーザ参加方法を用意することの有効性と，ユーザ知識を反映することで抽出精度が向上することを確認した.

今後の課題としては，開催日情報や開催地情報以外の抽出，人工知能学会のメーリングリスト以外のドキュメントへのシステムの適用，精度向上に寄与の大きいユーザ知識を，低い負荷で投稿できる方法の実現などがある.

参考文献

[1] Xin Xin, Li Juanzi, Tang Jie, and Luo Qiong. Academic conference homepage understanding using constrained hierarchical conditional random fields. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pp. 1301–1310, New York, NY, USA, 2008. ACM.

[2] 佐藤円, 佐藤理史, 篠田陽一. 電子ニュースのダイジェスト自動生成. 情報処理学会論文誌, Vol. 36, No. 10, pp. 2371–2379, 1995.

[3] 山田寛康, 工藤拓, 松本裕治. Support vector machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2002.

[4] 渡部啓吾, Danushka Bollegala, 松尾豊, 石塚満. Web からの人物の属性情報抽出. 第 23 回人工知能学会全国大会 (JSAI2009) 論文集, 2009.

[5] 後藤真孝, 緒方淳, 江渡浩一郎. Podcastle: ユーザ貢献により性能が向上する音声情報検索システム. 人工知能学会論文誌, Vol. 25, No. 1, pp. 104–113, 2010.