3H3-04

テキストからのメタデータ情報抽出とKnowWho検索への応用

Extracting Metadata from Text and Application to KnowWho.

岩倉 友哉 Iwakura Tomoya　塚本 浩司 Tsukamoto Koji　井形 伸之 Igata Nobuyuki　津田 宏 Tsuda Hiroshi

*1株式会社 富士通研究所 言語処理研究部
Language Processing Lab. Fujitsu Laboratories Ltd.

In this paper, we present our RDF metadata extraction technology from text and its application to KnowWho. Our metadata extraction consists of two steps. First, Named Entity(NE) extraction extracts entity types, such as "PERSON", "ORGANIZATION" and so on. Second, relation extraction extracts words that specify relations among NEs, such as PERSON's relation "friend" by syntactic pattern matching. As an application of our metadata extraction technique, we also show an automated KnowWho DB from news articles.

1. はじめに

SemanticWeb の実現に向けて，データについてのデータ (メタデータ) への感心が高まってきている．メタデータを利用することで，様々なアプリケーションが作成・統合できると期待されるが，利用するメタデータ整備のコストが問題となる．

我々は，スケジュール情報や報告書などグループウェア上の日常業務情報から人脈やスキルなどの人物情報メタデータを自動獲得する KnowWho システム「ヒューマンナレッジナビゲータ」[9, 3] を開発した．KnowWho とは，必要な知識を持った人を検索することである.「業務の情報の 50〜75%は人から直接得ている」というガートナー社の報告 [11] や，Google CEO の Eric Shmidt の"One of the problems with search is you can't find people." (2004.3.22) というコメントが示すように，今後，文書や KnowHow の検索から人を検索する KnowWho が重要になると予想される．

ヒューマンナレッジナビゲータが獲得しているメタデータ情報は，スケジュール共参加履歴から得られる人と人との関係の強さ，および作成文書中のキーワード頻度に基づく人とスキル情報の関係の強さである．関係の強さは有用な情報であるが，量的な指標だけでは十分に表現できず，質的な関係が有用な場面も存在する．たとえば,「岩倉」と「津田」との結びつきも，「岩倉」と「新人 B」との結び付きも，どれだけ同じミーティングに参加したかといった量的な指標にしてしまえば同じに見えてしまうが，実際には「部下」と「上司」や「新人同志」という関係があり，人と人との関係がわかることで，人脈はより活用できると考えられる.「人とスキル情報の関係」についても同様であり,「岩倉」と「言語処理」であれば関係は「専門家」,「岩倉」と「中国語」の関係は「学習者」のように，関係がわかることで，さらに適格な人を検索することができる．

このような関係情報を全て人手で書きつくすことは大きなコストがかかる．履歴書や報告書，社内文書，新聞や Web ページなどのテキスト中には関係情報が記述されており，テキストから関係付きメタデータを抽出することでメタデータ構築コストを下げることができると考えている．

本稿では，関係付きメタデータをテキストから自動で RDF 形式として抽出する手法の提案と，新聞記事で実験を行なった結果を報告する．

連絡先: 〒 211-8588 神奈川県川崎市中原区上小田中 4-1-1

表 1: 名詞のクラスの例

ARTIFACT	LOCATION	ORGANIZATION	PERSON
国民栄誉賞	日本	富士通研究所	太郎

表 2: 数値表現の例

DATE	MONEY	PERCENT	TIME
6月4日	５０００円	100 %	午前 10 時 10 分

2. 先行研究

International Semantic Web Conference 2003 で「自然言語 (テキスト) 処理による SemanticWeb」というテーマでワークショップ [13] が開催されるなど，SemanticWeb のためのテキストからのメタデータ情報抽出研究への期待は大きい．

テキストからのメタデータ抽出には，自然言語処理分野で研究されている Information Extraction(IE) 技術が広く利用されている．IE とはテキストから特定の情報を抽出する技術である．その中の Named Entity(NE) 抽出では,「⟨PERSON⟩岩倉⟨/PERSON⟩が⟨DATE⟩2004 年 6 月 4 日⟨/DATE⟩に発表する.」のように，テキストから，表 1 にある名詞のクラスや表 2 にある数値表現を抽出する．

[2] は，Web から，芸術家の人物情報を獲得するという研究を行なっている．構文解析，主語・目的語の解析，NE 抽出などの処理を行ない，WordNet で定義されている情報を利用して NE 間の関係抽出を行なう．たとえば，"Renoir was born on 25/2/1841." という文から，⟨PERSON⟩Renoir⟨/PERSON⟩は主語，⟨DATE⟩25/2/1841⟨DATE⟩は目的語，といった解析が行い，WordNet にある「動詞 bear の主語と目的語が持つ関係は，date_of_birth あるいは place_of_birth」といった定義から，解析結果を手掛りに適切な関係を選択し，

```
Renoir date_of_birth 25/2/1841
```

のような 3 つ組を抽出する．この手法では，WordNet に定義されていない関係は抽出できないという問題がある．

[7] は，人工知能学会の発表者間の "共著"，"研究室"，"プロジェクト"，"発表"，の関係を抽出することを行なっている．関係抽出は，発表者の名前がでてくるページを Web より集め,「ページ内で同じ行にでてくる」などを手がかりに教師あり学習手法で行なう．ここでの関係抽出は，特定ドメインのヒューリスティックスに基いているため，他の関係を抽出する場合は，新たに抽出の手掛りを用意する必要があると考えられる．

KnowWho に関係しては，[1] で，Web ページから人名抽出

を行ない，検索キーワードにマッチしたページから，そのキーワードに関連する人脈を獲得するという研究を行なっている．この方法では，検索キーワードごとに関連する人の継がりを調べることはできるが，人と人の関係の種類が明確でないという問題がある．

3. RDF メタデータ抽出のアプローチ

SemanticWeb では，Resource Description Format(RDF) がメタデータ記述言語として定義されている．我々のメタデータ抽出は，RDF 形式のメタデータをテキストから自動で獲得し，アプリケーションで利用することを目的としている．

RDF とは，(Subject， Predicate， Object) の 3 つ組を表現するものであり，これらの関係は，「Subject has a Predicate whose value is Object.」と解釈する．「http://xxx/ の作成者は岩倉.」という情報を RDF で記述する場合は，http://xxx/ を Subject，岩倉を Object，作成者を Predicate とし，Subject，Predicate，Object の順に，

〈http://xxx/〉〈#作成者 〉〈#岩倉 〉.

のような 3 つ組で記述する．実際には，Subject，Predicate，Object は，Uniform Resource Locator(URI) による記述が必要である．

我々は，テキストからの RDF 抽出を次の方法で行なうことを考えた．

- **Subject · Object 表現同定**：NE 抽出技術を用いて，テキストから ARTIFACT や PERSON といったクラスの名詞を抽出し，Subject，Object の候補とする．たとえば，「⟨ARTIFACT⟩ http://xxx/ ⟨/ARTIFACT⟩ の作成者は ⟨PERSON⟩ 岩倉 ⟨/PERSON⟩ .」のようなクラス付与により，Subject，Object を特定する．
- **Predicate 表現同定**：関係表現抽出により，Predicate に相当する表現を特定する．上の例であれば，関係を表す「作成者」を Predicate 表現として同定する．関係表現抽出方法としては，関係情報が記述された辞書の利用，構文パタンから関係表現を抽出する方法を用いる．
- **URI 付与**：テキストから RDF を生成するためには，人名の同姓同名のように表記が同じであっても実体は異なる場合や，表記のゆれ・省略表記・別名のように表記が異なっても同一であるという場合に対応しなければならない．このように，表記によらず，Subject， Predicate，Object の同一性を判定する必要がある．たとえば，「岩倉友哉」 は，「岩倉」と名字だけ記述される場合があり，これらが同じであるかを判別して URI を付与する．また，関係表現の場合は，既存のスキーマとのマッピングが必要となる．たとえば，Predicate 表現として抽出された「作成者」を Dublin Core の Creator にマッピングするなどである．

辞書を使って関係抽出を行なう手法に加え，構文パタンから関係表現を抽出することで，様々な関係が抽出できると期待される．現在は，NE 抽出ツールと関係表現抽出ツールの開発を中心に開発を行なっている．

4. 実装

本節では，4.1 節で教師あり学習手法による NE 抽出ツールの実装，4.2 節で関係表現抽出の実装について述べる．

4.1 NE 抽出ツールの実装

NE 抽出に関しては，IREX[4] などのコンテストが開催されており，多くの研究報告がある．開発手法としては，ルールベース，教師あり学習手法の二つが広く用いられている．教師あり学習手法 [8] は，正解データを用意することで，新しい領域や新規項目に対して適応できるという特徴を持つ．我々もこの点を考慮して，教師あり学習手法を採用することにした．

表 3: 形態素と NE タグの例

単語	品詞	NE タグ
神奈川	名詞-固有名詞-地域-一般	B-LOCATION
県	名詞-接尾-地域	I-LOCATION
川崎	名詞-固有名詞-地域-一般	I-LOCATION
市	名詞-接尾-地域	E-LOCATION

NE は複数の形態素で構成される場合がある．たとえば，「⟨LOCATION⟩ 神奈川県川崎市 ⟨/LOCATION⟩ 」であれば，「神奈川」，「県」，「川崎」，「市」という 4 つの形態素で構成される．教師あり学習手法による NE 抽出ツール開発においては，「形態素への “NE のクラス” と “NE 内での位置” のタグ付問題」として行なわれるのが一般的である．“NE のクラス” とは，PERSON，LOCATION などを指す．“NE 内での位置” とは，表 3 のように，それぞれの形態素が，NE 内で，先頭 (B-)，中間 (I-)，終わり (E-)，どの位置に属するかということを示すものである．実際には，図 3 の B-LOCATION のように NE の位置 と NE の種類 を組み合わせたタグを，現在位置および前後の形態素情報手がかりに判別することが一般的に行なわれている．形態素情報としては，単語の表記，品詞，単語の文字種などを利用する．

NE 抽出ツールは，[5] を参考に実装した．今回の実装では，単語の文字種定義の細分化を行ない，単語の最初と最後の一文字を素性として追加している．

教師あり学習手法は，一般的に性能が良いと言われている，Support Vector Machines(SVM) と，boosting を用いることにした．SVM を使った NE 抽出ツールの実装には，YamCha[*1] を用いた．boosting 学習器は，富士通研究所で開発した decision stump を弱学習器とする boosting 学習器 [6] を用いて行なった．

形態素解析器に ChaSen 2.3.3 [*2] を用いて，CRL データ[*3]から学習を行い，IREX の総合課題タスクで評価を行なった結果の Recall，Precision，F 値を表 4，5 に記す．IREX 総合課題とは，表 1, 2 にある 8 種類の NE を新聞記事より抽出するタスクである．2 値分類器である SVM を多値分類問題へ拡張するためには，各クラスごとに分類器を用意する one-vs-rest 法を用いた．カーネル関数は 2 次の多項式を利用した結果である．2 値分類器である boosting の多値分類問題への拡張は，[10] の手法で実装している．boosting の方が精度は劣る．しかし，簡単な速度比較を行なったところ，boosting が SVM より約 10 倍の抽出速度であった．多項式カーネルでの SVM の高速化は，[5] において提案されているが，decision stump を弱学習器とする boosting は，選択された素性へのインデックスの改良といった実装レベルでの高速化が有効に働くという点では利点があるといえる．

*1 http://chasen.aist-nara.ac.jp/ taku/software/yamcha/
*2 http://chasen.aist-nara.ac.jp/hiki/ChaSen/
*3 IREX において定義されている表 1, 2 の固有表現を，毎日新聞 95 年 1 月 1 日〜10 日までの全記事，約 1 万文に対して、タグ付けしたデータ．CRL データは付与された固有表現情報のみが IREX コンテストの総合課題の固有表現情報などとともにとも に http://www.csl.sony.co.jp/person/sekine/IREX/NE/ で配布されている．これらのデータは毎日新聞 94 年度，95 年度の CD-ROM があれば復元し利用することができる．

表 4: NE 抽出ツール (SVM) の IREX 総合課題での結果

NE	Recall	Precision	F 値
ARTIFACT	40.81	50	44.94
DATE	95.66	98.14	96.89
LOCATION	79.32	86.61	82.81
MONEY	100	93.75	96.77
ORGANIZATION	73.26	82.84	77.76
PERCENT	85.71	94.73	90
PERSON	88.16	89.42	88.79
TIME	96.61	95	95.79
全体	82.41	88.04	85.13

表 5: NE 抽出ツール (boosting) の IREX 総合課題での結果

NE	Recall	Precision	F 値
ARTIFACT	32.65	25	28.31
DATE	91.33	95.83	93.53
LOCATION	77.64	85.22	81.25
MONEY	100	100	100
ORGANIZATION	69.40	77.14	73.07
PERCENT	85.71	94.73	90
PERSON	85.35	87.57	86.44
TIME	94.91	96.55	95.72
Total	79.31	83.87	81.53

4.2 関係表現抽出ツールの実装

[2] にあるように，WordNet のような関係が記述された辞書を利用することで関係抽出を行なうことができる．しかし，人間関係のように数多くの関係種が存在するものを対象とする場合，全ての関係種とその表現を列挙した辞書を用意することは困難である．今回は，関係表現抽出のアプローチとして，構文パタンから NE 間の関係表現を抽出する手法を採用した．

現在，関係表現抽出ツールはルールベースにより開発しており，NE 間の関係表現が出現する構文パタンを，図 1 のようなルールとして作成している．ルール作成の手掛りとしては，形態素解析結果，NE 抽出結果を使っている．関係表現抽出ツールの実装は，形態素列に対する正規表現ツールを開発し，その上に関係表現抽出ルールを作成した．今回開発した形態素列に対する正規表現ツールでは，文字列の正規表現による処理のように，形態素列に対し，繰り返し・グルーピング・後方参照といった処理が行なえる．

たとえば，図 1 の構文パタンルールを用いることで，「⟨PERSON⟩ 岩倉 ⟨/PERSON⟩ の上司は ⟨PERSON⟩ 津田 ⟨/PERSON⟩ である．」という文から，$⟨ 関係語 ⟩ にマッチする「上司」が関係表現として抽出される．「⟨PERSON⟩ 岩倉 ⟨/PERSON⟩ は ⟨ORGANIZATION⟩ 富士通研 ⟨/ORGANIZATION⟩ の従業員．」という文からは，$⟨ 関係語 ⟩ にマッチする「従業員」が関係表現として抽出される．

このように，構文パタンを利用することで，辞書がなくとも NE 間の様々な関係表現を獲得できるという利点がある．さらに，辞書による関係抽出手法と組み合わせて利用することで，より多くの関係を抽出できると期待される．

5. KnowWho 検索への応用

NE 抽出ツール，関係表現抽出ツールを使ってテキストから人物に関するメタデータ情報を獲得することで，我々の KnowWho 検索を拡張できる．図 2, 3 は，1991 年〜2001 年の毎日新聞から表 1，2 にある 8 種類の NE 抽出を行ない[*4]，「ビートルズ」という単語が出現する 1146 記事から，「⟨PERSON⟩ ジョン・レノン ⟨/PERSON⟩ 」を中心に，PERSON(☆) と ORGANIZATION(※) を関連付けし視覚化した結果である．図 2，3 には，関連付け結果の上位ノードを視覚化したものを掲載している．視覚化部分は，[12] を利用した．

*4 NE 抽出ツールは，処理速度の面から boosting によるものを利用した．

```
$⟨PERSON⟩ の$⟨ 関係語 ⟩ は$⟨PERSON⟩
    → ($1 $2 $3)
$⟨PERSON⟩ は$⟨ORGANIZATION⟩ の$⟨ 関係語 ⟩
    → ($2 $3 $1)
```

図 1: 関係表現抽出ルールのイメージ

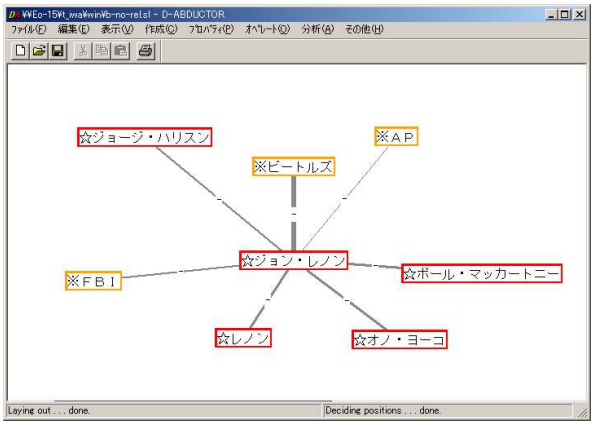

図 2: 毎日新聞から自動獲得したジョン・レノンの人物情報

図 2 は，従来のヒューマンナレッジナビゲータと同様に，「同じ文で共起したかどうか」で，関係の重みを計算することにより関連付けを行なった．この図より，「☆ジョン・レノン」は，「☆オノ・ヨーコ」，「☆ポール・マッカートニー」，「☆ジョージ・ハリスン」といった人物と継がりがあることがわかる．しかし，共起情報による関連付けのため，写真提供元として記事に出現する「※ＡＰ」との関連付けが行なわれている．また，「※ＦＢＩ」と「☆ジョン・レノン」の関係など，個々の関係はこの図からは読みとることはできない．

図 3 は，関係表現抽出結果で関連付けし，視覚化したものである．関係表現抽出により，「☆ポール・マッカートニー」，「☆ジョージ・ハリスン」がビートルズの「一員」，「メンバー」であることがわかる．また，「☆ジョン・レノン」の「親友」として関連付けされている「☆スチュアート・サトクリフ」など，共起情報が少なくて得ることができなかった人物との関係がわかる．「※ＦＢＩ」が「☆ジョン・レノン」は，ベトナム戦争の反戦運動関係の記事より，“怖がった” と関係付けされている．

抽出誤りもいくつか見うけられる．「☆来日」が ORGANIZATION として抽出されている．また，「伝記」，「クリスマスソング」が，「☆ジョン・レノン」と「※ビートルズ」の関係表現として抽出されている．図 2 では，「☆ジョン・レノン」，「☆レノン」，図 3 では，「☆ジョージ・ハリスン」，「☆ジョージ」と同一の人物が複数回マップに出現しており，同一性判定が必要であることがわかる．また，図 3 の「☆ポール・マッカートニー」と「☆ジョージ・ハリスン」の「※ビートルズ」との関係は，「メンバー」，「一員」となっており，関係表現の同一性判定も必要であることがわかる．

6. まとめ

本稿では，NE 抽出と関係表現抽出によるテキストからメタデータ情報獲得方法について述べ，KnowWho への応用について紹介した．NE 抽出技術と関係表現抽出技術により，テキ

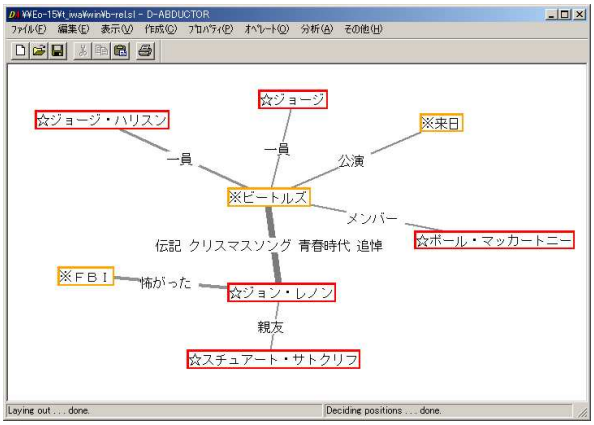

図 3: 毎日新聞から自動獲得したジョン・レノンの関係付き人物情報

ストから様々なメタデータ情報が低コストで獲得できると期待される．

メタデータ自動抽出においては，新たに抽出するクラスへの適応がいかに容易かが重要となる．教師あり学習手法では，IE などの知識がなくとも，抽出するクラスの定義に沿って正解を作成することで，抽出器が自動で適用するという利点がある．今回，関係表現抽出ツールはルールベースにより実装したが，NE 抽出ツールと同様に，教師あり学習手法による関係表現抽出を実現する必要があると考えている．

RDF の自動抽出に向けては，同一性判定，既存のスキーマへのマッピング技術の開発が次の課題である．

謝辞

NE 抽出実験にあたり，奈良先端科学技術大学松本研究室で開発されました茶筌，YamCha を使用させていただき，感謝申しあげます．

参考文献

[1] 原田昌紀，佐藤進也，風間一洋．Web 上のキーパーソンの発見と関係の可視化．情報処理学会情報学基礎研究会 2003.5.

[2] Harith Alani, Sanghee Kim, David E. Millard, Mark J.Weal, Paul H.Lewis, Wendy Hall, and Nigel Shadbolt. Automatic Extraction of Knowledge from Web Documents. Workshop on Human Language Technology for the Semantic Web and Web Services, 2003.10.20, Florida.

[3] 井形伸之，小櫻文彦，片山佳則，津田　宏．セマンティックグループウェア: RDF を用いた KnowWho の実現．人工知能学会 Semantic Web とオントロジー研究会, 2004.3.

[4] IREX 実行委員会（編）．IREX ワークショプ予稿集．1999.

[5] 磯崎秀樹，賀沢秀人．固有表現抽出のための SVM の高速化．情報処理学会論文誌，Vol. 44 No. 3 p970-979, Mar. 2003.

[6] Koji Tsukamoto, Yutaka Mitsuishi, and Manabu Sassano. Learning with Multiple Stacking for Named Entity Recognition. In Proceedings of the 6th Conference on Natural Language Learning,2002.

[7] 松尾豊，友部博教，橋田浩一，石塚満．Web からの人間関係ネットワークの抽出と情報支援．第 17 回人工知能学会全国大会，2003.6.

[8] 永田昌明．確率モデルによる自然言語処理．言語と心理の統計 ことばと行動の確率モデルによる分析．第二部．岩波書店．2003.

[9] Nobuyuki Igata, Hiroshi Tsuda, Yoshinori Katayama, and Fumihiko Kozakura. Semantic groupware and its application to KnowWho using RDF. ISWC(Intl. Semantic Web Conference) 2003 (poster), 2003.10.20-23, Florida.

[10] Robert E. Schapire and Yoram Singer. BoosTexter: A boosting-based system for text categorization. Machine Learning, 39(2/3):135-168, May/June 2000.

[11] The Knowledge Worker Investment Paradox. Gartner research, 2002.

[12] 渡部勇．富士通研究所による特許検索・分析支援システム「ACCENT」．INFOSTA2002, A-1, 2002.

[13] Workshop on Human Language Technology for the Semantic Web and Web Services. Proceedings of the ISWC 2003 Workshop, 2003.10.20, Florida.