文要素解析と固有表現抽出によるメタデータ抽出

SAITO, Yu / 齋藤, 悠

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編 / 法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
8
(開始ページ / Start Page)
65
(終了ページ / End Page)
68
(発行年 / Year)
2013-03
(URL)
https://doi.org/10.15002/00009570

文要素解析と固有表現抽出によるメタデータ抽出

Metadata Extraction by Sentence Element Analysis and Named Entity Extraction

齋藤　悠
Yu Saito
法政大学大学院情報科学研究科
E-mail: yu.saito.bg@stu.hosei.ac.jp

Abstract

There are a lot of documents without metadata about their semantic content on the web, but it is impracticable to manually add metadata to them considering costs; therefore a method to automatically extract metadata is necessary.

This paper proposes an approach to extract such 4W+HM metadata as <when>, <where>, <who>, <what>, and <hm> (stands for how-much and how-many) from Japanese plane texts and evaluates it to measure its retrieval effectiveness.

In this approach, metadata extraction process is free from dictionaries customized for specific fields, can apply general documents, and mainly consists of three parts: First, sentence element analysis part identifies what role each chunk in sentences plays; Second, named entity extraction part finds out remarkable expressions in sentences; Third, based on information given from previous parts, a criterion-based method outputs metadata.

The evaluation experiment of metadata extraction is performed by calculating recall and precision using a test set which is manually added correct metadata. Mextractr, which is related work to extract 5W1H metadata, is also evaluated with the same test set, and its results are compared with proposed method's. The experimental results show that proposed method is almost superior to Mextractr. In particular, proposed method excels in distinguishing <who> and <where>, which means that sentence elemental analysis works well not to confuse actors and places.

1. はじめに

Web 上の膨大な量の情報を効果的に共有・再利用するために，Web ページの構造のみならず意味内容を機械的に処理するための枠組みとしてセマンティック Web[1]がティム・バーナーズ＝リーによって提唱された．しかし，既存の Web ページの多くが人間に意味内容の解釈をゆだねる「電子的に書かれた読み物」に過ぎず，それらにセマンティック Web の基幹要素たるメタデータを人手によって付加するには労力や時間といったコストを要する．

そこで本研究では，セマンティック Web のデータソースを拡充するために，日本語のプレーンテキストから自動的にメタデータを抽出する手法を提案する．本研究では，情報の基本的な構成要素である 4W+HM（表 1 参照）メタデータを抽出対象とする．

2. 関連分野・研究

2.1. セマンティック Web

セマンティック Web とは，Web 上の情報を機械的に処理するための枠組みである．この枠組みでは，あくまで人間向けに記述された文書である Web ページに対し，文書の意味内容を補説する機械処理可能なメタデータを付与することで，プログラムによる情報の比較・統合を容易にする情報空間の実現を目指している．メタデータの語彙の差異はオントロジーと呼ばれる一種の辞書によって吸収される．

2.2. CNFE

Wang らが開発したシステム「Chinese News Fact Extractor (CNFE) [2] 」では，オンラインニュースの記事から 5W1H イベントに関する意味情報を抽出することを目的としている．CNFE では，まずニュースの見出しの重要性を強調するアルゴリズムによってトピックとなる文を抽出し，さらに抽出された文に対し，動詞に注目したルールと SVM による機械学習を併せた手法を適用することで 5W1H のイベント要素を抽出する．本研究と同様に文構造に注目した解析手法を用いているが，中国語文を対象にしている．

2.3. Mextractr

Mextractr[3]とは，メタデータ株式会社によって開発された商用メタデータ抽出エンジンである．Mextractr は，日本語で書かれたプレーンテキストの中から形態素解析，

表 1　メタデータの分類と対象例．

分類	対象例
when	絶対日時，相対日時
where	場所，住所
who	個人，企業，公共機関，委員会
what	商品，作品，イベント，法律，賞
hm (how-much /how-many)	金額，割合，個数，年齢，期間

Supervisor: Prof. Mina Akaishi

独自の知識ベース，要素別の判断アルゴリズムなどによって，意味属性の出現パターンやコンテクストを解析し，情報の基本的な要素である 5W1H の記述（いつ，どこ，誰，何，いくら，等）を自動的に抽出する．本研究と同じく日本語文を対象としていることから，評価実験にて提案手法と Mextractr の抽出結果を評価・比較する．詳しくは後述の評価実験の章を参照されたい．

3. 提案手法

現在のメタデータ抽出，ひいては情報抽出全体の技術は，対象となる分野特有の知識に依存する部分が多く[4]，新造語に対応するには知識が記述された辞書の更新が欠かせない．しかし，日々生まれ続ける新造語に追いつくには，人手によるメタデータ付与の問題と同様，コストがかかる．

そこで本研究では，対象に普遍的に含まれる情報，すなわち，文章ならば必ず存在する文要素と，文章中の特徴的な表現である固有表現を手がかりにすることで，企業名などの専用辞書の編纂や改訂を必要としないメタデータ抽出手法の提案をする．また，文要素に注目することで，地名姓（例：場所としての『秋田』と姓氏としての『秋田』）や，多義性（例：場所としての『日本』と国家行為体としての『日本（政府）』）などに起因する<where>と<who>の誤分類を低減する．

たとえば，<who>メタデータの抽出対象として銀行を考える．「三井住友銀行」や「三菱東京 UFJ 銀行」など「○○銀行」と名付ける慣例から普通名詞である接尾辞「銀行」を標識に抽出を行うと，「新銀行東京」といった銀行名には対応できない．網羅的な銀行名辞書を作成・利用することもできるが，そのような辞書を企業，政府機関，教育機関などあらゆる<who>メタデータの候補に用意することはコストの面で現実的ではなく，メタデータ自動抽出の本意に反する．

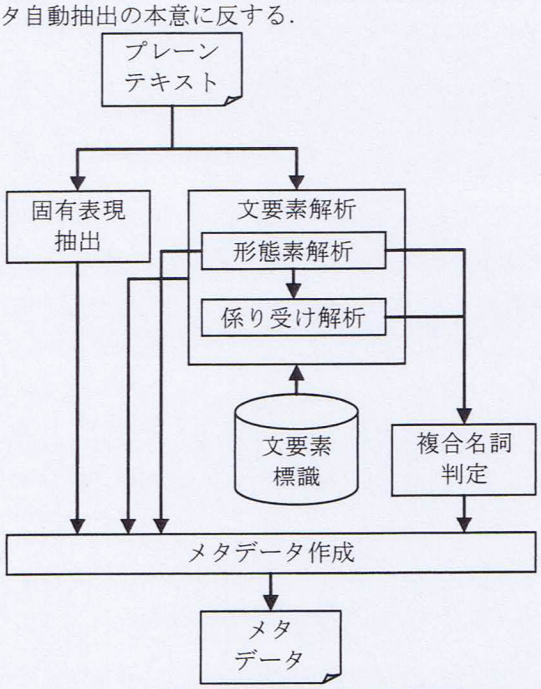

図１　メタデータ抽出プロセスの概略図．

そこで，図１に示すプロセスによってメタデータを抽出する．プレーンテキストは文要素解析と固有表現抽出に掛けられ，その処理結果を併せてメタデータ作成時の判定基準に利用する．文要素解析は形態素解析と係り受け解析，および文要素標識との比較によって行われ，中途の形態素解析の処理結果もまた個別にメタデータ作成時の判定基準に利用する．

3.1. 文要素解析

文要素解析とは，形態素解析，係り受け解析を経て，ある日本語文における各文節の機能的な分類を同定する処理である．機能的な分類とは，主語，目的語，補語[5]（状況語[6]）といった文要素を意味し，ゆえに文要素を同定するための標識を文要素標識と呼ぶ．文要素標識は格助詞，係助詞，副助詞，およびそれらの連語からなり，文要素とは表２に示す対応関係を持つ．

また，文要素解析は同時に並立構造の解消も伴う．並立構造とは「と」，「や」，「か」などの並立助詞によって並べられる，文中で同じ資格に立つ名詞の集まりである[5]．並立構造の解消では，並立助詞によって並べられた名詞の集まりを分解し，並列助詞によって省略された文要素標識に置き換える．

3.2. 固有表現抽出

固有表現抽出とは，テキスト中から固有表現を抽出する自然言語処理技術であるが，本研究では表３に示す IREX 定義に準拠した抽出を行う．

3.3. 複合名詞判定

複合名詞判定とは，形態素解析によって過分割された名詞を再結合する処理である．ひとつの文節に二回以上連続して名詞，あるいは未知語が登場した場合，それらの表記を連結し，複合名詞とする．

表２　文要素と標識．

文要素	標識
主語	～が，～は
目的語	～を，～に
補語 (状況語)	～へ，～で，～から，～まで，～までに，～への，～での，～からの，～までの

表３　IREX 定義の固有表現分類．

固有名詞的表現	組織名 政府組織名	ORGANIZATION
	人名	PERSON
	地名	LOCATION
	固有物名	ARTIFACT
時間表現	日付表現	DATE
	時間表現	TIME
数値表現	金額表現	MONEY
	割合表現	PERCENT

3.4 メタデータ作成

前段のすべてのプロセスの処理結果を使用して表４に示す判定基準によりメタデータを作成する．<hm>メタデータだけは「3000 件」のようにほぼ「数詞＋助数詞」の形をなしているため，形態素解析で得た品詞情報をそのまま流用する．また，「新銀行東京は中小企業に融資した」のように<LOCATION>東京</LOCATION>を含む複合名詞「新銀行東京」は主語であるので，<who>メタデータに分類される．このように，「新銀行東京」という語が辞書に登録されていない場合でも<who>メタデータとして抽出することが可能である．

表４　メタデータ作成時の判定基準．

分類	判定基準	例
when	単独の DATE 句	12月 19日，～
	単独の TIME 句	午前 8時ごろ，～
	単独の DATE＋TIME 句	12月 19日午前 8時ごろ，～
	補語∧DATE	今月 31日から～
	補語∧TIME	午後 3時までに～
	補語∧DATE＋TIME	今月 31日午後 3時までの～
where	単独の LOCATION 句	東京千代田区，～
	補語∧PER.を含む複合名詞	落合博満野球記念館で～
	補語∧ORGANIZATION	東京地方裁判所で～
	補語∧ORG.を含む複合名詞	ヤフードームで～
	補語∧LOCATION	東北への～
	補語∧LOC.を含む複合名詞	京都駅から～
	目的語∧LOCATION	尖閣諸島を～
	目的語∧LOC.を含む複合名詞	旭山動物園に～
who	単独の PERSON 句	～（本名、鈴木栄治）～
	主語∧PERSON	室伏広治が～
	主語∧PER.を含む複合名詞	伊藤忠商事は～
	主語∧ORGANIZATION	東京地方裁判所が～
	主語∧ORG.を含む複合名詞	小岩井農場は～
	主語∧LOC.を含む複合名詞	横浜 DeNA ベイスターズは～
what	目的語∧ARTIFACT	グッドデザイン賞を～
	目的語∧単独の未知語	iPad を～
hm	MONEY	10万円，20億ドル
	PERCENT	8割，12倍
	数詞＋助数詞	100戸，3000件

4. 評価実験

提案手法を実装し，メタデータ抽出結果の評価実験を行う．また，固有表現抽出の結果をそのまま 4W+HM に写像したものと，Mextractr の Web API 版を利用して得た結果との比較を行う．

4.1. 実装

評価実験を行うに当たり，提案手法を表５に示すツールを用いて実装した．新造語に対応できているか確認するために，CaboCha [7]は 2003 年作成の資源を使用している．

表５　実装に用いたツール．

開発環境	Eclipse 4.2
使用言語	Java
形態素解析器	JUMAN 7.0
係り受け解析器	KNP 4.01
固有表現抽出器	CaboCha 0.53

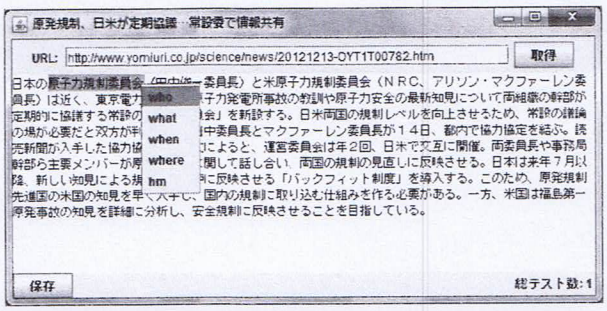

図２　テストセット作成支援プログラム作業画面．

4.2. テストセットの準備

提案手法と Mextractr のメタデータの抽出結果の正否を評価するに当たり，図２に示すテストセット作成支援プログラムを自作し，読売新聞が提供しているニュースサイト「YOMIURI ONLINE」のニュース記事からテストセットを作成した．計 103 のニュース記事の本文（プレーンテキスト）に対し，人手によって 4W+HM の正解メタデータが付与されている．

4.3. メタデータ抽出結果と考察

テストセットに対する提案手法の抽出結果を表６，Mextractr の抽出結果を表７，固有表現抽出のみの抽出結果を表８に示す．再現率および適合率はそれぞれ次式（１）および（２）で与えられ，再現率はどれだけ正解メタデータを抽出し損なっているかを表す指標であり，適合率はどれだけ余計なメタデータを抽出しているかを表す指標である．双方の指標とも値が大きいほど良い性能ということになる[4]．

$$\text{再現率} = \frac{\text{抽出された正解メタデータの数}}{\text{正解メタデータの総数}} \qquad (1)$$

$$\text{適合率} = \frac{\text{抽出された正解メタデータの数}}{\text{抽出されたメタデータの総数}} \qquad (2)$$

表６　提案手法の抽出結果．

	再現率	適合率
when	0.849	0.778
where	0.770	0.703
who	0.711	0.783
what	0.241	0.413
hm	0.880	0.947

表７　Mextractr の抽出結果.

	再現率	適合率
when	0.729	0.698
where	0.671	0.510
who	0.577	0.508
what	0.094	0.135
hm	0.160	0.915

表８　固有表現抽出のみの抽出結果.

	再現率	適合率
when	0.849	0.636
where	0.801	0.322
who	0.541	0.511
what	0.152	0.419
hm	0.232	0.951

押し並べて本研究による抽出結果の方が Mextractr の抽出結果を上回った．特に，<where>と<who>の再現率に大きな違いはないが，適合率では提案手法が上回っている．これは，Mextractr での抽出に起きる<where>と<who>の混同が原因と思われる．提案手法では，たとえば「東京地方裁判所で傍聴した」のような場所としての「東京地方裁判所」と，「東京地方裁判所は無罪判決を言い渡した」のような行為主体としての「東京地方裁判所」を明確に区別しているため，このような結果になったと思われる．

また，提案手法と固有表現抽出のみの抽出結果を比較すると，提案手法が適合率を底上げしているのが分かる．これは，固有表現抽出によって過分に集められたメタデータの候補が，文要素解析によって精錬されているためと思われる．

<what>の抽出結果が最も悪い結果となったのは，人工物という幅広い括りであることと，本研究の基本方針が特定分野の辞書群に依存しないことが起因したと思われる．

<hm>の抽出結果は提案手法も Mextractr も最も良い結果となった．これは，数量表現はパターンマッチングでほぼ網羅できるためであろう．

5. まとめ

本研究では，文要素解析と固有表現抽出を併用して，プレーンテキストから自動的に 4W+HM メタデータを抽出する手法について提案した．評価実験においては，既存手法よりも全体的に高い適合率および再現率を計測し，提案手法の有効性を示すことができた．

また，提案手法では表４に示す判断基準のみを挙げたが，他にも抽出に寄与する判定基準は存在し得る．

6. 今後の課題

抽出結果の改善につながる課題として，<what>メタデータに関しては異なるアプローチを用いることが挙げられる．新しい商品や新しい事件・事故など，時事性の高い<what>メタデータには Wikipedia と Twitter のような網羅性と即時性の高いデータソースを利用することが挙げられる．実際に，文献[8]では幅広い概念を取得するために WWW という広大な探索空間を解析していた従来の手法に対し，Wikipedia という一サイトの閉じたリンク構造を利用することで現実的な解析時間を実現し，新造語への対応を困難にしてきた「辞書データを作成するときと利用するときのタイムラグ」を，Wikipedia の記事がリアルタイムで更新され続けるという点に着目し，解消を図っている．また，文献[9]では Twitter 上の出現頻度が急上昇した注目語を固有名詞として取得する手法について取り組んでおり，形態素解析を正しく行うことが困難な商品や作品に関する言及の解析精度向上を図っている．

また，提案手法で用いた文要素解析は，あくまで表層に現れる標識をもとに行われるもので，深層構造を考慮したものではない．ゆえに，主語にも目的語にも付くため双方の区別が難しい「～も」，「～だけ」，「～ばかり」などは標識から除外されており，それらの標識を伴ってしか言及されないメタデータは抽出できない．対象となるプレーンテキストの文章が平易ではなく，複雑な変形が極端に含まれていると，抽出できるメタデータ数は減少してしまう．この課題を解決するには，深層構造推定といった意味解析に踏み込むアプローチが必要であろう．

文　献

[1] Semantic Web - W3C: http://www.w3.org/standards/semanticweb/

[2] W. Wang, D. Zhao, L. Zou, D. Wang, and W. Zheng, “Extracting 5W1H Event Semantic Elements from Chinese Online News,” Web-Age Information Management Lecture Notes in Computer Science, vol.6184, pp.644-655, 2010.

[3] 5W1H メタデータ自動抽出ソフトウェア Mextractr : http://www.mextractr.net/

[4] 徳永健伸，情報検索と言語処理，辻井潤一（編），（財）東京大学出版会，東京，1999.

[5] 高橋太郎，日本語の文法，（社）ひつじ書房，東京，2005.

[6] 鈴木重幸，日本語文法・形態論，（社）麦書房，東京，1978.

[7] 工藤拓，松本裕治．“チャンキングの段階適用による日本語係り受け解析，”情報処理学会論文誌，vol.43，no.6，pp.1834-1842，2002.

[8] 中山浩太郎，原隆浩，西尾章治郎．“Wikipedia マイニングによるシソーラス辞書の構築手法，”情報処理学会論文誌，vol.47，no.10，pp.2917-2928，2006.

[9] 加藤慶一，秋岡明香，村岡洋一，山名早人．“ミニブログにおける注目語抽出手法の提案と注目語を用いたメディア間での話題追跡，” 情報処理学会研究報告，vol.2010-DBS-151，no.22，pp.1-8，2010.