

Web 検索と EC サイト, Wikipedia を利用した固有物名の抽出

荒井 徹[†] 大和田 勇人[‡]

東京理科大学理工学研究科経営工学専攻[†]

東京理科大学理工学部経営工学科[‡]

1. はじめに

近年, 情報化が浸透し, 個人が入手できる情報量は膨大なものとなっている. 情報の形態としては, テキストや画像, 音声など様々な種類が存在するが, 中心として用いられているのはHTML やワープロ文書に代表されるテキストデータである. インターネット上に存在するテキストデータを全て読むことによって必要な情報を探し出すことは事実上不可能である. そこで膨大な量の出来データからユーザの要求にあったものを見つけ出す情報検索技術に加えて, 大量の文書を構造化・組織化する技術が重要視され研究されている. それらの技術を用いることでユーザに必要な情報を絞り込むことが可能になると考えられている.

文書の構造化の手法として, 近年では情報抽出が注目を集めている. 情報抽出とは, テキストデータを有効活用するため, 特定目的の情報を抽出してデータベースなどに入力することで, テキストデータを構造化する技術である. その中で基礎技術となるのが固有表現抽出である. 固有表現とは, 固有名詞 (組織名・人名・地名・固有物名), 時間表現 (日付・時刻), 数値表現 (金額・割合) といったものを指し, これらを高精度に抽出することが重要となる.

ここで固有物名以外の固有名詞や時間表現, 数値表現といったものは既存研究で9割以上の精度を持つものがある一方で, 固有物名に注目すると5割程度の精度しかなく, まだ十分でないと言える. そこで本論文では, 固有物名の抽出に着目し, 固有物名を漏れなく抽出できる手法を提案する. そこで使用するのはWeb 検索である. Web 上のテキスト量は教師データの典型的なサイズに比べて桁違いに大きく, 言語資源として非常に利用価値の高いと考えられる.

2. 固有表現抽出

固有表現抽出は新聞記事などのテキストデータに対し, テキスト中のどこからどこまでが固有表現部分なのかを判断し, その種類は何かを判定するタスクである.

固有表現抽出は様々な研究機関によって研究されている. ヨーロッパ系の言語を対象とした固有表現抽出はMUC (Message Understanding Conference) や CoNLL (Conference on Natural Language Learning) によって盛んに研究され, 日本語を対象とした固有表現抽出はIREX (Information Retrieval and Extraction Exercise) においてさまざまな手法が比較されている. そして固有表現である固有名詞・時間表現・数値表現の各々がどこまでの表現を含むかについては曖昧な面がある. そのため上記のような評価会において正解判定をするために厳密な定義が必要である.

IREX では, 固有表現抽出システムはテキスト中に存在する固有表現文字列の開始・終了位置に重複や入れ子の無い唯一のタグのペアをふり, もし, 表現が重なっている場合は, 原則的に長い単位の表現を抽出する, としている. 例えば, 「日本銀行」は「日本」を地名として抽出するのではなく, 「日本銀行」全体を組織名として抽出する. IREX では, 固有名詞 (組織名・人名・地名・固有物名), 時間表現 (日付表現・時刻表現), 数値表現 (金額表現・割合表現) を抽出対象としている. 以下にIREXで定義された固有表現を説明する.

- 組織名 <ORGANIZATION>: 複数の人間で構成され, 共通の目的を持った組織などの名称を指す. 株式会社などの会社, 固有の政府組織, 学校, 国際組織やなんらかの目的を持ったグループなども組織としての意味で使われている文脈においては組織名とする
- 人名 <PERSON>: 固有の人物を表す名称を指す. 役職名, 敬称などは人名に含まない.
- 地名 <LOCATION>: 固有の場所を表す名称を指す. 国名, 都道府県名, 市町村名などの行政単位となりうるものから, 河川名, 山脈名などの地形に関するものを含む. ただし, 地方, 地域, 周辺, 内, 圏, 諸国, 方角, 部, 沿岸, 沖などのついた概略的表現は地名表現には入れない. 例えば, 「関東地方」は「関東」のみを地名として抽出する.
- 固有物名 <ARTIFACT>: 人間の活動によって作られた具体物, 抽象物を含む物の名前を指す.
- 日付表現 <DATE>: 特定の時間を表現するもので, その単位が24 時間以上のものであるものを指す.
- 時間表現 <TIME>: 特定の時間を表現するもので,

The extraction of ARTIFACT that used Web search and EC site, Wikipedia

[†] Toru Arai · Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science

[‡] Hayato Ohwada · Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

その単位が24時間以下のものであるものを指す。

- 金額表現 <MONEY> : 金額を表す表現。
- 割合表現 <PERCENT> : 割合を表す表現。

固有表現抽出は、辞書を用意するだけで簡単に解決しように思えるが、実際には辞書を作るだけでは不十分であり、前後の文脈を含めた判断が必要になることが多い。また固有表現は多種多様であり、次々と新しい表現が生み出されるため、そのすべてを辞書に登録することは不可能である。

3. 提案手法

ここでは固有物名を漏れなく抽出するために、まず固有名詞を全て抽出し、その後固有名詞の中で組織名・人名・地名・固有物名といった種類に分類していく。

1. まず、与えられた文書に対して形態素解析を行い、どこからどこまでが何という単語であるかを把握し、文書を形態素ごとに分割する。
2. 次に、フレーズでWeb検索を行う。IREXの定義において原則的に長い単位の表現を抽出するとなっているため、長い文節から順に検索を行っていく。そして一定以上の割合で検索数が増加した場合、固有名詞である可能性が高い判断する
3. その後、新たにWikipediaと指定したECサイトで検索をし、そこで検索されれば固有名詞であると断定する。
4. そして検索されたページを解析し、固有表現が組織名・人名・地名であるかを判定し、それ以外のものを固有物名とする。

この手法を採用することで、今まで抽出出来ていなかった固有物名を抽出することが出来る。

4. 評価実験

今回、提案手法の評価を行うために、インターネット上のニュース記事やブログ記事などから固有物名を500語含むように文書を集めた。そして手動でラベル付けを行うことで精度を評価した。この時、しきい値を検索数が1桁以上または1000以上増加したときとした。結果を以下の表1に示す。

表1. 提案手法の精度

適合率	再現率	F-measure
0.948	0.71601208459	0.815834768

表1から分かるように適合率が9割を超え、今回集めた文書中の固有物名のほとんどを抽出することが出来、良い精度となっている。一方で固有物名ではないものも固有物名として抽出してしまっているため、再現率が7割程

度となっている。そのため総合的な精度であるF-measureが8割程度となっている。

5. まとめ

本論文では既存研究では十分な精度がなかった固有物名に着目し、Web検索を用いて固有名詞を抽出する手法を提案した。提案手法では従来手法での精度が低かった固有物名を漏れなく抽出することで、固有物名の抽出精度を高められた。しかし評価実験の結果からF-measureが9割を超えていないので、まだ実用段階にあるとは言えず、さらに精度を上げていかなければならない。

また、この提案手法では固有物名を漏れなく抽出するために、長い文節から順に検索を行っていく。そのため、文書全体から形態素を少しずつ少なくして検索するという手法をとっているため、処理数が莫大な数になってしまうという問題点がある。この問題点は重大なものであると考えられるので、今後はこの処理を少なくしていく必要がある。

また、提案手法では固有名詞を漏れなく抽出するようにしているため、イベント名や事件名も抽出してしまっている。それらは基本的に他の固有表現よりも単位が長い場合、間違っ抽出してしまう。そのため、イベント名や事件名を除外するようにしていくことで再現率を上げていくことが課題となる。

参考文献

- [1] IREX (NE) Home Page, <http://nlp.cs.nyu.edu/irex/NE/>
- [2] 竹本義美, 福島俊一, 山田洋志. 辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出. 情報処理学会論文誌, Vol. 42, No. 6, pp. 1580-1591, 2001.
- [3] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 形態素解析システム『茶筌』version2.2.9 使用説明書. 奈良先端科学技術大学院大学, 2002.
- [4] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 165(1):91-134, 2005
- [5] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142-147. Edmonton, Canada.
- [6] B. Settles, "Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets," in *NLPBA/BioNLP*, 2004.
- [7] G. Zhou and J. Su. Named entity tagging using an HMM-based chunk tagger. In *Proceedings of the 40th Annual meeting of the ACL*, pages 209-219, Philadelphia, PA, July 2002.