自然言語処理技術の現状と展望
エラー分析プロジェクトを通して

基礎 一般

3.4 固有表現抽出

岩倉 友哉（(株) 富士通研究所）

固有表現抽出とは

固有表現とは，人名（PERSON）や組織名（ORGANIZATION）などの固有名詞や，日付（DATE）や時間（TIME）などの数値表現を指し，テキストから固有表現を抽出する処理を固有表現抽出と呼ぶ．たとえば，次のような抽出を行う．

<PERSON> 太郎 </PERSON> さんの誕生日は <DATE>8 月 4 日 </DATE> です．

固有表現抽出は，テキストからの情報抽出のための要素技術の 1 つとして，Message Understanding Conference-6（MUC-6）[1] において定義された．固有表現の種類は目的に応じて異なる．経営トップの交替などの情報を抽出することを目的としていた MUC-6 では，LOCATION，ORGANIZATION，PERSON，DATE，MONEY，PERCENT，TIME の 7 種類であった．その後，日本語を対象とした Information Retrieval and Extraction Exercise[2] においては，ARTIFACT（法律名や製品名などの人工物）が加えられた．そのほかには，約 200 種類で構成される拡張固有表現 [3] が提案されるなど，固有表現の種類は広がりを見せている．また，用途の面においても，質問応答における解答候補の獲得，テキスト中の個人情報の匿名化など，幅広く使われるようになってきた．

抽出手法概要

固有表現抽出では，同じ表記の単語であっても文脈によって意味が変わる場合を区別して抽出する必要がある．たとえば，以下の文からは，「宮崎」を PERSON と LOCATION として抽出する．

<PERSON> 宮崎 </PERSON> さんに
<LOCATION> 宮崎 </LOCATION> で会う．

固有表現を抽出する手法の 1 つとして，人名や地名の辞書を用いる方法がある．しかし，辞書との照合による抽出では，上記の例の「宮崎」を PERSON および LOCATION と区別して抽出できない．また，辞書に含まれていない固有表現は抽出できないという問題もある．そこで，多くの場合は，辞書だけでなく，文脈情報も手掛かりとして用いる．以下に，機械学習に基づく固有表現抽出手法の概要を紹介する．

この方法は，固有表現タグ付きコーパスを用意することで，抽出用のモデル・規則を自動的に獲得できる．また，コーパスサイズを大きくすることで精度改善が行えるため，近年主流となっている．機械学習による固有表現抽出の実現手法の 1 つとして，各単語に固有表現のラベルを付与する分類器を学習することが挙げられる．この方法では，「<PERSON> 宮崎 </PERSON> さん」のように，抽出したい個所に固有表現のタグを付与した学習用コーパスを用意する．続いて，日本語であれば，学習用コーパスの各文から，形態素解析器で単語を切り出し，「宮崎 /PERSON さん /O」のように各単語に対応する固有表現タグを元に正解ラベルを付与する．「O」は固有表現以外の単語という意味である．抽出の手掛かりである素性としては，対象単語やその前後の単語の表記や品詞，単語と辞書との照合結果などが用いられる．機械学習を用いることで，たとえば，「“さん” が直後に出現する単語は PERSON の可能性が高い」といった学習が行われる．

固有表現抽出の課題

BCCWJ☆1 コーパスに含まれる新聞やブログなどの 6 分野の文書を用いて，京都大学黒橋・河原研究室にて開発されている KNP☆2 の固有表現抽出機能のエラー分析を行った．そこから洗い出した課

題を以下に紹介する．

学習用コーパス・辞書の問題

機械学習に基づく固有表現抽出においては，学習用コーパスが精度に関係してくる．BCCWJ を用いた調査では，学習データに出現した固有表現の正解率は，出現しなかった固有表現の正解率と比較し，LOCATION で 40 ポイント以上，PERSON で 20 ポイント以上高いという結果であった．また，KNP が学習データとして用いた新聞記事では，再現率と適合率の調和平均である F-measure が約 83 であったが，文体の異なるブログや雑誌などの文書では，F-measure は 60 前後であった．

また，辞書のカバレッジも精度に大きく影響する．KNP が用いている形態素解析器 JUMAN☆3 の辞書に登録されている LOCATION，ORGANIZATION，PERSON の正解率は，JUMAN の辞書に含まれてない固有表現の正解率と比較し，30 ～ 40 ポイントほど高いという結果であった．さらに，評価に用いた文書中に出現するすべての LOCATION，ORGANIZATION，PERSON を形態素解析器の辞書に登録することで，それぞれの F-measure が 10 ポイント以上改善した．

このように学習用コーパスや辞書の精度に対する影響は大きい．今後は，人手による言語資源の整備に加え，学習用コーパスや辞書の自動獲得手法のさらなる発展が望まれる．

曖昧な語への対処の問題

固有表現抽出では，「宮崎」のように文脈によって PERSON や LOCATION と異なる固有表現となる語や，「ライオン」のように，普通名詞にも固有表現にもなり得る曖昧な語の意味を区別する必要がある．現況の多くの抽出手法では，これらの意味を区別するために，対象単語の前後数単語や，それらの品詞，辞書との照合結果といった局所的な文脈情報を主な手掛かりとするのが一般的である．そのため，辞書に登録されており，学習用コーパスに出現している場合であっても，曖昧な語は，抽出器が対象とする文脈の範囲に十分な情報が含まれないと，正しく判別できないことが多い．たとえば，

クマには命を助けられたことがある．

という文では，「クマ」が固有表現か判別するために，同文書内のほかの文を参照する必要がある．この例では，「森で野生のクマに会った．」という文が前文にあれば固有表現ではなく，「クマさんこと，篠原氏の小説．」といった文があれば，PERSON として抽出するのが正しい．今後，このような例に対処するために，1 つの文だけでなく，前後の文など，より広い文脈情報を考慮する手法が必要である．

常識や意味の問題

抽出のために語の実体の知識が必要になる場合がある．

バンプレストさんのソフトで，仮面ライダーやウルトラマン，ガンダムが 2 頭身で一緒になって戦うソフトの名前なんでしたっけ？？

この例では，「さん」を手掛かりとした場合，ORGANIZATION である「バンプレスト」を PERSON として抽出してしまう．正しく抽出するためには，「仮面ライダー」や「ウルトラマン」に関する「ソフト」を持つ会社が「バンプレスト」という知識が必要だと考えられる．今後，このような問題に対処していくためには，エンティティリンキングといった，語の実体を判別する技術との組合せが挙げられる．

参考文献

1) Grishman, R. and Sundheim, B. : Message Understanding Conference-6 : A Brief History, In Proceedings of the 16th Conference on Computational linguistics (1996).
2) Sekine, S. and Isahara, H. : IREX : IR and IE Evaluation Project in Japanese, In Proceedings of the Second International Conference on Language Resources and Evaluation (2000).
3) Sekine, S., Sudo, K. and Nobata, C. : Extended Named Entity Hierarchy, In Proceedings of the Third International Conference on Language Resources and Evaluation (2002).

(2015 年 11 月 2 日受付)

岩倉 友哉（正会員）iwakura.tomoya@jp.fujitsu.com
2003 年（株）富士通研究所入社．2011 年東京工業大学大学院総合理工学研究科物理情報システム専攻博士課程修了．博士（工学）．現在，（株）富士通研究所主任研究員．自然言語処理の研究開発に従事．

☆1 http://pj.ninjal.ac.jp/corpus_center/bccwj/
☆2 http://nlp.ist.i.kyoto-u.ac.jp/?KNP
☆3 http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN