日本語、英語テキストからの固有名・数量情報抽出システム

デモ8

下畑 光夫　福本 淳一
沖電気工業(株) 研究開発本部 関西総合研究所

1 はじめに

インターネットの普及などにより、電子データが急増している現在、情報抽出に対する期待が高まっている。我々は、固有名・数量情報の抽出を情報抽出の基礎技術と考え、開発を行っている。その一環として98年4月に行われた情報抽出の国際会議であるMUC-7、MET-2に参加し、MUC-7(英語テキスト対象)では、適合率92%、再現率77%を、MET-2(日本語テキスト対象)では、適合率97%、再現率85%を達成した[1]。

固有名・数量情報は、テキストに記述されている情報の重要かつ基礎的要素であり、様々な分野でこれらの情報を利用することができる。特に固有名の属性は、その固有名の果たす機能、役割を限定することから広く用いることができる。以下、2節で抽出方式について述べ、3節でこの抽出方式を利用した応用事例について述べる。

2 抽出手法の概要

固有名は種類が非常に多い上に組織名などは新しく生成されることもあり、固有名を網羅するような辞書の作成、更新は非常に困難である。固有名は表層的な構成に特徴があることが多く、この表層的構成を利用して認識することで、未知の固有名も含めて膨大な固有名を認識することができる。

また、語単独では様々な属性をとりうる固有名が多く存在するために辞書で静的に属性を決定しても適合しない場合が多い。属性決定においては、固有名特有の接辞を優先したり、既抽出の固有名を利用することで、文脈に応じた属性を付与している。また、数量表現は、「数字列＋数量単位語」を基に抽出している[2]。

Some Applications of Named Entity Recognition
Mitsuo Shimohata, Jun'ichi Fukumoto
Kansai Laboratory, R&D Group, Oki Electric Industry Co., Ltd.

2.1 日本語テキスト用固有名抽出

日本語の固有名では様々な属性をとりうる語が多く存在することから、辞書的情報より接辞を優先して属性を決定している。例えば、「地名＋“社長”」という箇所では、辞書的情報と接辞でコンフリクトが生じるが、接辞を優先して地名にあたる語の属性は人名に転化される。また、新聞記事では、人名に必ずなんらかの接辞を付与することから接辞のない単独の固有名(“千葉”、“山口”など)は地名としている。拘束力の強い接辞では、前接語が未知語であっても固有名と認識している。例えば、「カタカナ文字列＋(さん|氏)」という文字列では、カタカナ文字列が未知語であっても人名として抽出する。

また、固有名内の語の構成により大半の組織名を抽出している。例えば、銀行の固有名を認識する場合、「地名＋“銀行”→固有銀行名」という抽出規則により多くの銀行が抽出できる。この方式は、組織名辞書のサイズをコンパクトにするとともに、未知の組織名を抽出できるという長所がある。

2.2 英語テキスト用固有名抽出

英語では、固有名はキャピタライズされているため、キャピタライズされている語が連結した領域を切り出すことで固有名領域を抽出することができる。ただし、文頭の語は固有名でなくてもキャピタライズされているので、文頭語処理を行っている。固有名領域中の先頭または最後の語から属性を判別している。人名であれば“Mr.”や“President”といった固有名特有の語により属性を決定することができる。

また、英語の新聞記事では、固有名が最初に出現する時はフルネームや接辞を伴って表記されるが、2度目以降に出現する固有名は略称で参照されることが多く、固有名単独での属性決定が困難である。そこで、複数の語からなる抽出固有名の構成語を以降の固有名抽出に利用している。例えば、記事中に「Dr. Washington」

という表記があれば、その記事で「Washington」と単独で出現した場合、人名と判定している。

3 固有名・数量情報抽出の応用事例

2節で述べたように、本抽出方法は属性を含めて固有名を抽出できることや、日英両言語のテキストから抽出することができることから様々な応用が考えられる。以下にいくつかの応用事例について述べる。

抽出情報のマークアップ

固有名・数量情報抽出単独の利用事例として、WWWブラウザに抽出モジュールを組み込み、HTMLテキスト中の固有名に属性に応じたマークアップを行なっている。これにより、各固有名の視認性を高めている。構文情報を用いずに抽出処理を行なっているため、処理が軽く、HTMLデータの読み込みと抽出処理をリアルタイムで行なうことができる。

イベントによる記事情報抽出

テキストからドメイン特有のイベントを含む文をキーとして情報抽出を行なうことにより、テキストの中心的情報を抽出することができる。ドメインを新製品情報にし、ドメイン特有の「発売する」や「発表する」といったイベントを含む文から固有名と数量表現を抽出することで、新製品に関する情報 (名称、発売日、価格など) を抽出している。

固有名間の関係情報抽出

2つの固有名がなす関係 (part_of や kind_of など) は、検索システムにおいて検索語の関連語拡張で利用されるなど有用性が高い。本固有名抽出方式では、固有名の属性とその間と直後の助詞を用いることで固有名間の関係を認識している。

「AのB」といった書き方で表される助詞「の」の前後に現れる2つの名詞は様々な関係をとることができるが、前後の名詞の属性が判明している場合には高い確率でその関係を一つに限定することができる。例えば、「"組織" の "人名"」という表記であれば、ほとんどの場合人名と組織は所属関係にある。また、直後に「を」「は」「に」などの係り受けの流れを止める確率が高い助詞があることを条件にすることで精度を上げている。

対訳コーパスからの固有名辞書作成

本抽出方式は、日本語と英語の両言語に対して同程度の抽出精度を実現している [3]。対訳テキストから固有名を抽出し、属性の等しい固有名で対応付けを取ることで固有名辞書を作成することができる。固有名が未知語であっても、接辞により属性を判定することができるため、未知語でも固有名辞書も作成することが可能となっている。

統計的手法により対訳辞書を作成する方法 [4] でも未知語に対する固有名辞書を作成することができるが、統計的信頼性を満たすために多くの対訳テキストが必要となる。このような辞書作成方法は少量のコーパスからでも信頼性の高い辞書を作成することができると考えられる。

4 まとめと今後の課題

本論文では、情報抽出の要素技術である固有名と数量情報の抽出方式について、また抽出した情報を応用した事例について述べ、様々な分野に適用できることを示した。

今後は、固有名抽出方式の改良を行なうと共に、固有名抽出の応用事例を引き続いて検討していく予定である。また、参照関係や構文解析などを導入し、より高度な情報抽出も目指していくつもりである。

参考文献

[1] *Proceedings of Seventh Message Understanding Conference* (1998).

[2] 下畑, 福本, 杉尾：パターンと構文情報による固有名の情報抽出, 言語処理学会第4回年次大会ワークショップ論文集, pp. 44–49 (1998).

[3] 福本, 下畑, 桝井：固有名詞抽出における日本語と英語の比較, 情報処理学会自然言語処理研究会 98-NL-126, pp. 107–114 (1998).

[4] 北村, 松本：対訳コーパスを利用した対訳表現の自動抽出, 情報処理学会論文誌, Vol. 38, No. 4, pp. 727–736 (1997).