日本語文書に含まれる固有用語の自動抽出方式

3J－5

奥　雅博　　東田　正信

ＮＴＴ情報通信処理研究所

１．はじめに

機械翻訳やキーワード自動抽出などの日本語解析技術を応用した日本語処理システムの開発が盛んに行われている。これらのシステムでは，システムが持つ電子化辞書を用いて入力文を正確に解析することが必要である。従って，入力される文書に対して，辞書がどの程度充実しているかがシステムの精度を決める要因の１つとなる。特に，適用分野を考えると，文書の種類（新聞，論文など）などによって辞書に登録すべき語が異なる。

適用分野に適合した辞書を作成するためには，一般語以外に辞書に登録すべき語（固有名詞や専門用語など）が数多く存在する。本稿では，このような語を『固有用語』と呼び，次のように定義する。

《『固有用語』の定義》

ある日本語文書でのみ使用されている製品名，会社名，人名などの固有名詞や，既知の単語の組み合わせからなる語であっても全体として新語（未知語）や専門用語，および，その文書中においてのみ使用されていると考えられる語。

このような固有用語を文書中から自動的に抽出することは，キーワードの抽出や機械翻訳における辞書の充実などを支援する技術として重要である。

２．固有用語抽出方式

従来，固有用語の抽出方法としては，

・日本語解析に基づく方法[1)]，

・表層情報を利用した方法[2)]，

が提案されている。前者の方法は，日本語解析において解析不能となった部分（未知語）を抽出する。しかし，この方法では，本来１語として扱われるべき図１に示すような固有用語をその構成単語に分割することによって解析に成功してしまい，固有用語を抽出することができないという問題がある。これに対して，後者の方法では，この問題を避けることができるが，辞書登録語などの固有用語ではないものも抽出してしまうという問題がある。

固有／運動　：　proper motion
固有／Ｘ線　：　characteristic X-rays
固有／関数　：　eigenfunction

＜図１：　固有用語の例＞

そこで本稿では，表層情報をもとに文書中から固有用語の候補を抽出した後，一般語辞書を引くこと，およびルールを適用することの２つの処理によって，固有用語として適切な用語のみを抽出する方式を提案する。本方式は以下の３つの処理からなる。

（ａ）字種の変化点に着目した候補の抽出，

（ｂ）辞書引きによる候補の絞り込み，

（ｃ）ルールによる候補の絞り込み。

（ａ）字種の変化点に着目した抽出

処理（ａ）では，日本語の持つ豊富な字種に着目し，字種の変化点から固有用語の候補を抽出する。字種は，抽出の精度を上げるために，従来よりも細かく，表１に示す９種類とした。

表１：　字種の種類

項番	コードの名前	項番	コードの名前
1	漢字ｺｰﾄﾞ	6	ｱﾙﾌｧﾍﾞｯﾄｺｰﾄﾞ
2	漢数字ｺｰﾄﾞ	7	句読点ｺｰﾄﾞ
3	ひらがなｺｰﾄﾞ	8	区切りｺｰﾄﾞ
4	ｶﾀｶﾅｺｰﾄﾞ	9	その他のｺｰﾄﾞ
5	算用数字ｺｰﾄﾞ		

固有用語の候補として次の文字列を抽出する。

《ひらがなｺｰﾄﾞ，句読点ｺｰﾄﾞ，区切りｺｰﾄﾞ
のいずれかの字種に挟まれている文字列》　…①

①に従って抽出される候補の例を図２に示す。

ＰチャンネルＭＥＳＦＥＴを用いると遷移動作は低速になり，ガリウム・ひ素を用いた相補型論理回路に対してはシリコンを…

▒▒▒▒　…　固有用語の候補

＜図２：　固有用語の候補抽出の例＞

（ｂ）辞書引きによる候補の絞り込み

処理（ａ）では，既知／未知に関係なく，①を満たす文字列を固有用語の候補として抽出する。このため，辞書登録語も抽出されてしまう（例えば，図２「低速」）。処理（ｂ）ではこれらに対して辞書検索を行うことによって既知の語を候補から削除し，固有用語の候補を絞る。

（ｃ）ルールによる候補の絞り込み

処理（ｂ）までで得られた固有用語の候補には，固有用語ではないが，接辞を含むために処理（ｂ）で削除されなかった文字列（例：新会社）や，図２中の「用」のような単語の部分文字列が含まれている。これらの削除

A Method for the Extraction of Proper Words in Japanese Texts
Masahiro OKU, Masanobu HIGASHIDA
NTT Communications and Information Processing Laboratories

すべき文字列は，これを構成する字種の組み合わせによってその特徴が異なる。この異なりに着目して，文字列を字種の組み合わせに応じて分類し，分類ごとに削除すべき候補の文字列の持つ特徴をルール化する。そしてこのルールを満足する候補は削除すべき文字列であるとして候補から削除する。表２に本稿で用いた文字列の分類を示し，表３に分類ごとに作成したルールの一部を示す。

表２：　文字列の分類

分類番号	対象となる文字列の条件
分類　１	漢（数）字１文字
分類　２	漢（数）字２文字
：	：
分類１５	ｱﾙﾌｧﾍﾞｯﾄ列に算用数字列が連続
：	：
分類２１	算用数字列
分類２２	上記以外の文字列

表３：　ルールの例

分類	条　　件	処　　理
分類１	無条件	候補から削除する
分類２	接辞を含む	候補から削除する
分類３	接辞，または 副詞性名詞を含む	候補から削除する
：	：	：

処理（ｃ）では，このルールを用いて，固有用語の候補のうち，それが属する文字列の分類に記述されているルールを満足する文字列を候補から削除する処理を行う。

３．机上評価実験

３．１　実験の概略

本稿で提案した固有用語抽出方式の有効性を検証するために机上での評価実験を行った。処理（ｂ）における辞書としては，「三省堂新明解国語辞典」を用い，処理（ｃ）に関しては，新聞記事から作成したルールを用いてそれぞれ人手により処理を行った。対象文書としては，ルール作成に用いた新聞記事２１６文の他に論文８０２文，高校教科書２１９文，マニュアル２１７文を用いた。

有効性の評価は，この実験で抽出された固有用語と，定義に従って人手により抽出した固有用語とが，どの程度一致しているかによって行い，一致の度合いを示す指標として再現率，適合率の２つの値を用いた。

・再現率　＝　Ｃ／Ｂ　（１－検出もれ率）
・適合率　＝　Ｃ／Ａ　（１－誤抽出率）
　Ａ：　本方式により抽出された固有用語数
　Ｂ：　人手により抽出された固有用語数
　Ｃ：　Ａのうち，Ｂに含まれる数

再現率は抽出すべき固有用語のうち，何割が抽出されているかを示し，適合率は抽出されているもののうち，何割が正しく抽出されているかを示す。

３．２　実験結果と考察

実験結果を表４に示す。再現率は約９５％以上の高率を得ており，適合率についても６６％～８１％の値が得られている。人手抽出を支援するという立場に立つと，少々の誤抽出があっても抽出もれがないことが望まれる。この観点から見ると，表４の再現率，適合率はかなり良い値であり，本方式は固有用語の自動抽出法として有効であると考えられる。

表４：　評価実験結果

	再現率［％］	適合率［％］
新聞記事	９６．６	８３．９
論　文	９４．８	６６．０
教科書	９７．８	８１．１
マニュアル	９９．３	７２．６

・再現率について

本方式は原理的にひらがなを含む固有用語を抽出できない（例えば図２「ガリウム・ひ素」）。特に論文では，連用形名詞を含む固有用語（例：係り受け解析）が多く，再現率の低下の一因となっている。再現率を高めるためには，処理（ａ）の段階において，ひらがなを含む固有用語を候補として抽出する必要がある。

・適合率について

適合率は，ルール作成に用いた新聞記事に対して８４％，他分野に対して６６～８２％となっており，ルールの分野依存性が認められる。このことから，分野ごとにきめ細かなルールを作成することにより，適合率を向上させることができると考えられる。

４．おわりに

本稿では，日本語文書から固有用語を自動的に抽出する方式について提案し，机上実験を通してその有効性を検証した。今後，抽出精度をさらに向上させるために，

・処理（ａ）において抽出する文字列の条件として，字種の変化点だけでなく，字種の並び方についても考慮することによって，ひらがなを含む固有用語を抽出する手法，
・より複雑なルール（候補の一部分を取り出して固有用語とするなど）を記述するための枠組み，

の２点について検討を進める。

【参考文献】

（１）新井：翻訳支援システムＴＥＲ－ＧＥＴの概要，情処　自然言語処理研究会　87-NL-62　(1987).
（２）高橋，渥美，高橋：マニュアルの索引と用語集の作成支援，第３７回情処全大　5B-3(1988).