5C-8 知的情報検索システムＩＲＩＳにおける固有名詞抽出用形態素解析

川崎　正博　　伊吹　潤　　秋山　幸司
（富　士　通　株　式　会　社）

1．はじめに

日本語質問文を理解し，回答となる内容を持つテキスト群をテキストベースから検索する知的情報検索システム IRIS ［杉山(1986)］では，現在，対象分野を情報産業界の新聞記事見出しとして，システムの拡張および，実験を行っている。テキストが新聞記事見出しという性格上，未登録会社名，未登録製品名等の固有名詞の出現頻度が高く，大規模テキストベースにおいては，辞書整備の工数も膨大なものになるため，それら未登録固有名詞に対応できる処理が必須になった。また，その固有名詞がどういう意味クラスに属する（会社名であるとか製品名である等）かも推測できるとより良い文解析が見込まれることになる。本稿では，会社名と製品名を中心とした未登録固有名詞の切出しとその固有名詞の意味クラスの推定を行う固有名詞抽出処理のアルゴリズムおよび実験結果，今後の課題を述べる。

2．全体の構成

固有名詞抽出処理は，図１のように，まず，与えられた入力文より，固有名詞と思われる部分文字列のみを切り出し，次に，ヒューリスティクスに基づいてその固有名詞の意味クラスを推定し，出力するものである。

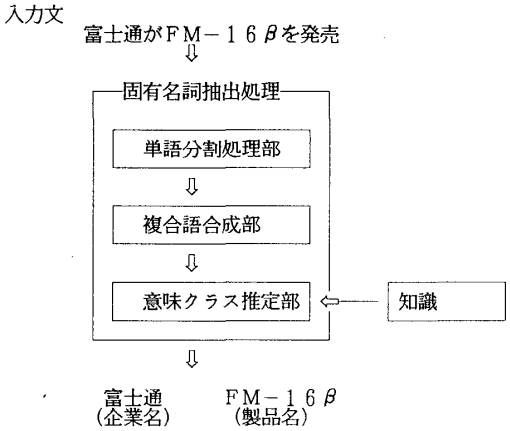


図１　固有名詞抽出処理の概要

固有名詞抽出処理の全体の構成は，入力された文字列を辞書を用いて分割する単語分割処理部，名詞が連続している部分を合成し，ひとつの単語とする複合語合成部，切り出された文字列（固有名詞候補）の意味クラスを推定する意味クラス推定処理部に分れる。以下におのおのの処理部について述べる。

3．単語分割処理部

アルゴリズムとしては，最適探索手法を用いた分節数最小法を使った。入力はかな漢字混じり表記の新聞記事見出し文である。この処理部の特徴としては，漢字，ひらがな，カタカナなどの字種を利用し未登録固有名詞候補を取り出すことである。その動作は，字種が変化したある文字位置より１つ以上の名詞と１つ以上の未登録語が連続して検索された場合に限り，字種が次に変化する位置までを未登録固有名詞候補として取り出すことである。例えば，入力文が“富士通山形が・・・・”という場合においては，図２のように“富士”“通”“山形”と同じ字種（漢字）のなかで３つの単語が検索される。しかし，それらの単語のうちでいずれかひとつの単語が未登録語であった場合に限り，次の字種の変化位置までを切り出した“富士通山形”を含むパス２が出力とされる。これは，固有名詞が未登録語を含んだ複数の単語候補に分割されるのを防いだためである。固有名詞が複数の登録語に分割される場合においては，次の複合語合成部で対処している。

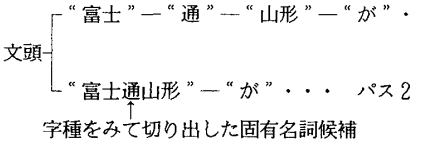


図２　未登録固有名詞を含んだ入力文字列の解析例

4．複合語合成部

この処理部では，おのおのの単語が持つ情報および前後の助詞情報をもとに，既登録名詞が連続している部分をひとつにまとめる処理を行う（勿論，おのおのの単語が持つ情報は保存しておく）。これは，固有名詞が複数の登録語によって分割されるのを防ぐために行うもので，例としては図３－ａのようなものがある。図３－ｂのように，固有名詞でないものまで，ひとつにまとめてしまう可能性があるが，それらは，次の意味クラス推定処理部で失敗し，排除される。

5．意味クラス推定部と知識

この処理部では，名詞連続合成部において出力された文字列の意味クラスの推定を行う。会社名，製品名がどのようなパターンで用いられる場合が多いかを調査し，そのうえで，会社名知識パターン

Morphological Analysis for the recognition of proper noun in IRIS

Kawasaki Masahiro, Jun Ibuki, Kohji Akiyama
Fujitsu Limited

（表1），製品名知識パターン（表2）を作成した。文字列がそのパターンに一致したものであれば，それぞれの意味クラスを与える処理を行う。一致しないものは，未登録会社名，未登録製品名でないと思い，排除する。

・入力文
　科学技術情報センターが・・・・・
・単語分割処理部出力
　“科学”—“技術”—“情報”—“センター”—“が”・・
・複合語合成部出力
　“科学技術情報センター”—“が”・・
　（a）

”紫外線”—”機器”・・・・
⇩
”紫外線機器”・・・・
（b）

図3　名詞連続合成部の役割り

5．1　会社名パターン知識

情報処理に関連する新聞記事の見出しより，数百の会社名を取り出し，パターンを調べた結果，以下の3つのパターンにあてはまるものが多いことが判った。

① アルフベット3文字において成り立っているもの
② 文字列の先頭あるいは，最終単語に地名を示す単語が含まれているもの（日本，東京が多い）。
③ 会社名の語尾に用いられる頻度が高い単語で終っているもの。

表1　会社名知識パターン

パターン	例
① 3文字ですべてアルファベット	ＩＢＭ，ＮＴＴ，ＡＴＴ　など
② 文字列の先頭あるいは終端に地名を表す単語がくる。	日本ＩＢＭ，日本電線，東海エンジニアリング，富士通山形　など
③ 文字列の終端に会社名に良く用いられる単語がくる。	○○電機，○○銀行，○○商事，○○運輸，○○工業，○○通信，○○電気，○○証券，○○保険，○○工業　○○建設　など

5．2　製品名パターン知識

会社名パターン知識のように明確なものが少なく，今回の実験では，次のパターンにあてはまるものだけを製品名固有名詞とすることにした。

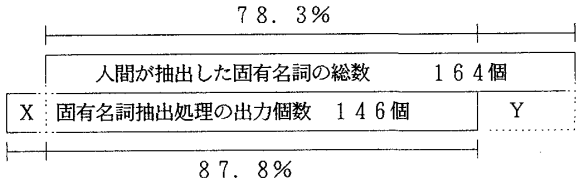
① その文字列がアルファベット（長さ3文字以外）のみで成り立つかあるいは数字を含んだアルファベット列で成り立つもの。

表2　製品名知識パターン

パターン	例
① アルファベット列である（数字も含む）。	ＰＣ－９８０１，ＦＭ－１６β　ＯＡＳＹＳ　など

6．評価

情報処理に関連する新聞記事の見出しを200 文，取り出し，実験を行なった。大見出し，中見出し等の区切りは◆で示した。知識として，企業名パターン知識の③のパターン数は３７個，辞書は，人名，地名を含んだ約１０万語を用いた。検索対象となる新聞記事見出し200 文のなかから，人間が固有名詞と思われるものをすべて抽出し(164個），固有名詞抽出処理の出力がどの程度，人間が抽出したものと一致するかを調べた結果を図4に示した。人間が抽出した固有名詞の総数をAとし，固有名詞抽出処理が出力した固有名詞の総数をB，そのうち，人間が抽出したものと一致した固有名詞の総数をCとした時，C／B（固有名詞抽出処理が出力したうちで人間が抽出したものと一致した割合）が87.8%，C／A（人間が抽出した固有名詞のうちで固有名詞抽出処理によって出力されたものの割合）が79.3% となる。なお，固有名詞として抽出されなかったものはカタカナ混じりの固有名詞がほとんどであり，その他は会社名が省略して記述された場合に人名，地名と同じ表記になってしまうもの（”松下”等）であった。



・固有名詞抽出処理において出力されたもののうちで固有名詞でないもの　X　１６個

・固有名詞であるのにもかかわらず，固有名詞抽出処理で抽出されなかったもの　Y　３４個

図4　固有名詞抽出処理の実験結果

今後は，固有名詞であるにもかかわらず抽出されなかったものについての検討を重ね，大規模なデータについて，少ないパターンでどれほどの固有名詞を抽出できるかを実験する予定である。

謝辞：本研究は第5世代コンピュータプロジェクトの一環として行われた。御支援頂いた ICOT の方々に深謝致します。

【参考文献】

[杉山(1986)]　杉山ほか「自然言語理解に基づく情報検索システムＩＲＩＳ」情報処理学会自然言語処理研究会資料58-8, 1986.