1S−2

英語形態素解析用辞書のコンパクト化

川崎　正博　松井くにお　西野　文人
(富士通研究所)

1．はじめに

一般的に，自然言語処理に用いられる単語辞書は十万語にのぼるような大規模なものが多く，一単語が持つ形態素情報，構文情報等のデータも，細かく表現されている反面，複雑化し，扱いにくいものとなっている事が多くみられる。しかし，実際においては，そのような大規模辞書をそのまま利用する事は少なく，システムに必要な情報のみを取り出したシステム用辞書を作り出し，辞書コストの低減を図っている事が多い。　本稿では，英語を入力文とし，品詞の推定等を行う形態素解析処理において，そのシステムの特徴を生かし，名詞類を品詞として持つ単語を辞書より削除することによる辞書のコンパクト化の実現方法，および，そのコンパクト辞書を用いて英語形態素解析（Ｅｍｏｒ）を行った実験結果，今後の課題を述べる。

2．英語形態素解析（Ｅｍｏｒ）

英語形態素解析Ｅｍｏｒは，英語入力文に含まれる単語の活用形処理，品詞推定，スペル検査・修正，未登録語処理を，行の長さや空白，辞書中に含まれる単語の頻度，品詞間の推移確率によって解析をおこなう統合的な英語形態素解析ツールである〔1〕。

3．コンパクト化処理の全体構成

まず，コンパクト化を行う処理の全体構成としては，図1のように，従来型辞書からの単語レコードを入力として受取り，その単語から，英語形態素解析システムに必要な情報のみを取り出す単語情報取り出し部，品詞情報によって，その単語レコードをコンパクト辞書に登録するか否かを決定する品詞判定処理部，登録条件が満たされた単語レコードをコンパクト辞書に登録する単語レコード登録部によって成り立つ。以下におのおのの処理部について詳細に述べることとする。

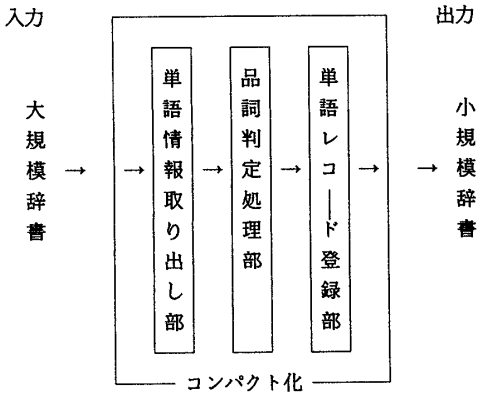

図1　コンパクト化処理の全体構成

A compact dictionary for use in part-of-speech analysis
Masahiro KAWASAKI, Kunio MATSUI, Fumihito NISHINO
FUJITSU LABORATORIES LIMITED

4．単語情報取り出し部とタグコード

英語形態素解析に必要な情報として，単語見出し，単語使用頻度，および，タグコードがある。それぞれの説明として，単語見出しは，その単語の持ちうる表層上の文字列，単語使用頻度は，その単語の用いられる割合を示し，タグコードというものは，品詞情報と形態素情報を組み合わせたもので，単なる名詞においても，固有名詞単数形（所有格），普通名詞単数形（所有格以外）などの８つの分類がなされており，動詞などにおいても，現在形，過去形，過去分詞形，進行形等に分類されている〔表1〕。名詞が分類されたタグコードを持つ単語を，ここでは，名詞類を品詞として持つ単語とする。この単語情報取り出し部においては，タグコード別に使用頻度が付与されている為，タグコードに対応づけて，単語見出しを与える（同じ見出しにおいても，タグコードが違えば，別レコードとする）。そのようなレコードを作成したうえで，次の品詞判定処理部の入力とする。

表1　タグコードの例

N	普通名詞単数形〔所有格以外〕	book
N.PL	普通名詞複数形〔所有格以外〕	books
N.PL.PS	普通名詞複数形〔所有格〕	books' (s)
N.PR	固有名詞単数形〔所有格以外〕	Smith
N.PR.PL	固有名詞複数形〔所有格以外〕	Smiths
N.PR.PL.PS	固有名詞複数形〔所有格〕	Smiths' (s)
N.PR.PS	固有名詞単数形〔所有格〕	Smith's
N.PS	普通名詞複数形〔所有格〕	book's
V	動詞現在形(3人称単数形以外)	
V.ED	動詞過去形	

・
・
・

5．品詞判定処理部

単語情報取り出し部から入力とされたレコードより，単語見出しの同じものに対して，そのすべてが，品詞として，名詞類を持つ時，その見出し語をコンパクト辞書に登録せずに，同一見出しで，名詞類を品詞として持つ単語レコードもあるが，他の品詞を持つ単語レコードもあるものや，すべてが名詞類を品詞として持たない単語レコードである場合は，その見出しを持つ単語レコードすべてを登録する。ここで，部分的に名詞類を持つレコードをなぜ削除しないかというと，その単語が解析結果として，他の品詞（例えば，動詞など）に推定され，出力する可能性を持つ事になる可能性があるからである。具体例として，表2で示すように，従来型辞書より，単語情報取り出し部で整形されたレコード（単語見出し　book, book stand, book store）がある。bookは名詞類と動詞類を品詞として持ち，book stand, book storeは名詞類の品詞のみしか持たない。その為に，コンパクト辞書に登録されるのは，bookのみ（図中における＊マークのもの）ということになる。

表2　単語情報取り出し部の出力例

見出し	タグコード
＊ book	N　（普通名詞単数形〔所有格以外〕）
＊ book	N.PS （普通名詞複数形〔所有格〕）
＊ book	N.PL （普通名詞複数形〔所有格以外〕）
＊ book	V　（動詞現在形）
＊ book	V.BD （動詞過去形）
＊ book	V.EN （動詞過去分詞形）
＊ book	V.ING （動詞ing 形）
＊ book	V.S （動詞現在形３人称単数形）
book case	N　（普通名詞単数形〔所有格以外〕）
book case	N.PS （普通名詞複数形〔所有格〕）
book case	N.PL （普通名詞複数形〔所有格以外〕）
book stand	N　（普通名詞単数形〔所有格以外〕）
book stand	N.PS （普通名詞複数形〔所有格〕）
book stand	N.PL （普通名詞複数形〔所有格以外〕）

6．単語登録レコード登録部

名詞類のみを品詞として持つ単語レコードを除いたレコード群に対し，ここの処理部では，辞書に登録するうえでの圧縮を行う。見出し語に対し，前方の文字列が同じになるようなもの（bookとboomに現れるような先頭３文字boo が同じもの）には，後に現れた単語に対して，前方に現れた単語と違う表記である部分（この場合 m）のみを表示し，あとは同じである部分の長さ（この場合　3）を示すようにする。そうする事によって，有効にコンパクト化が行われることとなる。

7．実験と実験結果

単語情報取り出し部の出力をそのまま辞書としたコンパクト辞書１と品詞判定処理部の出力をそのまま辞書としたコンパクト辞書２（名詞類を品詞として持つ単語を削除したもの）について，以下の順序で実験を行った。

a）　コンパクト化された辞書を作成する。

b）　コンパクト化された辞書において，形態素処理を行う。

c）　辞書の大きさ比較，形態素解析の異なり調査を行う。

コンパクト辞書１は従来型の大規模辞書から品詞推定処理に不必要な情報を取り除いただけのものであるので，形態素解析結果は大規模辞書をそのまま使用したものと同様である。コンパクト辞書２は，品詞判定処理部において，名詞類のみを品詞として持つものをすべて削除しているため，全体の大きさとしては，表３のようになり，大幅なコンパクト化が図られたことがわかる。なお，表中では，単語登録部の出力をそのまま辞書としてものをコンパクト辞書３とした。

表3　辞書の比較

辞書名	単語数〔語〕	バイト数〔byte〕
大規模辞書	198377	21411840
コンパクト辞書１	198377	3176917
コンパクト辞書２	66611	889136
コンパクト辞書３	66611	622877

なお，このコンパクト化による悪影響がどれほど解析結果に出るかが問題であるが，英語形態素解析システム（Ｅｍｏｒ）の品詞推定を実験の対象とし，英語の基本文956 文（単語6994個）を入力とした場合の結果を表４にまとめた。なお，コンパクト辞書２を辞書として用いた場合の未登録語はすべて名詞として扱ったうえでの比較とした。

全体的には違いが見られた文が187 文（全体の19.6％），単語250個（全体の約3.6％），そのうち，名詞熟語が分割されて出力されたもの123 個（違いが見られた単語のうちの49.2％）であった。その他のものとしては，ある名詞類を持つ品詞が違うものに推定されている21個（8.4 ％），動詞類を持つ品詞が違うものに推定されている39個（15.6％）などであり，名詞類のみを品詞として持つ単語を削除してもほとんど悪影響が出ない事がわかる。

表4　解析異なりの分類

分類	割合％	例
名詞熟語から	４７．５	rush-hour　rushとhourに分割
動詞類から	１５．６	stand(V)がstand(N)に
名詞類から	８．４	meat(N.PL)がmeat(N) に
副詞類から	８．４	when(ADV) がwhen(N) に
形容詞類から	４．８	public(ADJ) が未登録語名詞に
その他	１５．３	接頭語　be動詞など

8．今後の課題

実験結果にわかるとおり，名詞熟語（rush-hour,等）の異なりが多く現れている。しかし，これについては，決して誤りといえないものであるため，今後は，名詞熟語が分割された状態において，どれほど，誤りとされる単語となりうるか等の調査と，それ以外のものの異なりの検討を行ってゆき，原因を調査したいと思っている。今後の改良としては，辞書より名詞的概念を品詞として持つ単語のみでなく，lyを単語の語尾として持つ副詞を削除するなど，いろんな面での辞書のコンパクトを検討してゆきたいと思っている。

〔参考文献〕

〔1〕西野　”英語形態素解析Ｅｍｏｒにおける品詞推定”　情報処理学会第４１回全国大会予稿集　1990.