文節に対するコスト付け手法を用いた形態素解析システム

4AE－4

栄留孝行　兵藤安昭　若田光敏　池田尚志
岐阜大学工学部

1　はじめに

形態素解析において，解を1つに絞り込む為の方法として，コスト最小法があるが，従来行われてきた方法は，各々の単語と単語間にコストを付け，その値の合計が最小となるものを最適解として選択するものである．しかし，これでは，単語の2項関係しか見ないので，長い形態素の連なりとなると誤った解析結果が得られる事も多い．

そこで我々は，単語間にコストを付けるのではなく，大域的なコストを考慮した形態素解析システムを作成した．

文節に対するコスト付けに使用する文節コストデータベースは，新聞記事(155,000文)を対象に，計算機で形態素解析した結果から正解と思われる文節を抽出し，その出現頻度をもとに作成した．

2　システムの概要

形態素解析用の自立語辞書は，EDR日本語単語辞書をベースに，ひらがな以外の同一文字種からなる，名詞，サ変名詞を削除したもの(約21万語)を用いた．また機能語については，上記の辞書に加えて，複合的な機能語(約1800語)を登録してある．

システムの概要を図1に示す．まずシステム起動時には，形態素辞書をメモリに読み込む．辞書は，パトリシア形式のツリー上に登録し，検索は，パトリシア構造を利用した高速で効率的な最長一致検索を行っている．

辞書引き後，可能な文節候補すべてを作成し，これらの文節パターンに対する文節コストを付与する．文節パターンは，図2に示す規則に従って文節をパターン化したものである．

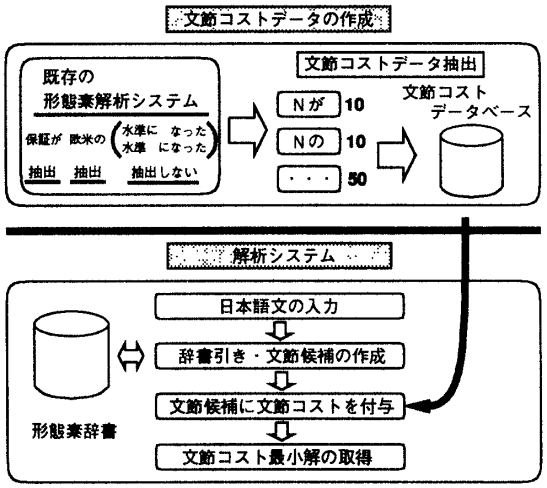

図1: システムの概要

コスト最小のパスはViterbiアルゴリズムによって見出す．

3　文節コストデータベース

文節コストデータベースは，我々が開発したヒューリスティックに基づく形態素解析システム[1]における解析結果を利用し，文節パターンの出現頻度を元に作成する．

この形態素解析システムは，3つの最長一致法アルゴリズム(単純な前方・後方最長一致法と，後方最長一致法に補正を加えた方法)の解を融合して，正解が含まれる可能性の高い解集合を求めるものである．この3つの最長一致法において同一の解が得られた文節は，正しく解析される可能性が高いと考え，文節コストデータ抽出の対象とした．

抽出された文節は，次のようにパターン化する．自立語は，図2に示すように12種に分類し登録する．機能語は，用言の直後に付属する活用語尾については省略し，その他は，表記のまま登録する．

抽出された文節パターンには，頻度を元にして10～500まで10段階のコストを付ける．

朝日新聞記事社説 155,792 文を対象に，上記の方法で，文節コストデータの抽出を行ない，1,862,456 文節から 23,549 種の文節パターンを得た(表1／図3)．

Morphological Analysis System using Bunsetsu Pattern Cost
Takayuki Eidome, Yasuaki Hyodo, Mitsutoshi Wakata, Takashi Ikeda
Faculty of Engineering, Gifu University
Gifu-shi,501-11,Japan

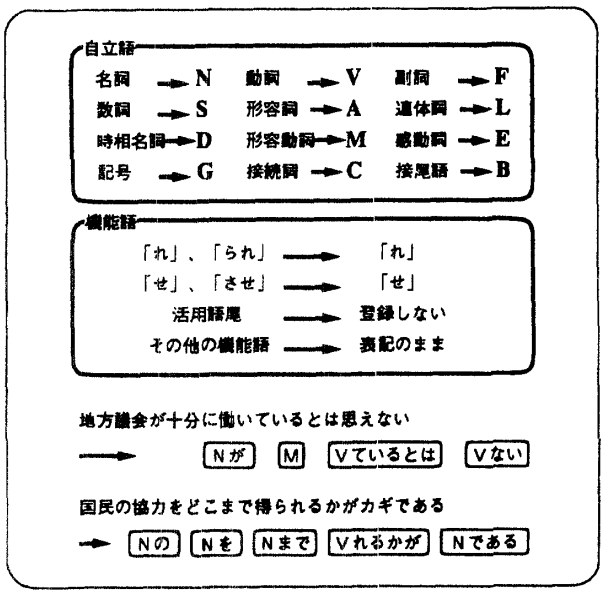

図 2: パターン化規則

パターン	頻度	cost	パターン	頻度	cost
Nの	75931	10	Vないよう	50	80
F	21965	20	Nかどうか	20	80
Vた	13168	30	Vつつあるという	5	200
Nでの	1112	60	Nではありません	2	200
Nをも	104	70	Nにさいしての	1	200

表 1: 文節パターンの例

この，文節コストデータの精度は，500 文 (6401 文節) を調査したところ，パターン化された文節の中で 98.4% が正しく解析されていた．

4　解析実験

文節コストデータベース作成に用いた社説と，それ以外の天声人語，各 100 文について評価実験を行った．成功例として以下のような長い単語列を 1 つの文節として切り出している．

『願わぬことはないが，』
解析結果：(願 / わ / ぬ / ことはな / い / が / 、)

解析に失敗した文の内訳を表 2に示す．

失敗内訳	社説	天声人語
未知語による失敗	2	6
品詞の選択に失敗	2	1
機能語・接続規則の不備	1	1
文節コストデータの不備	1	1

表 2: 解析失敗文の内訳

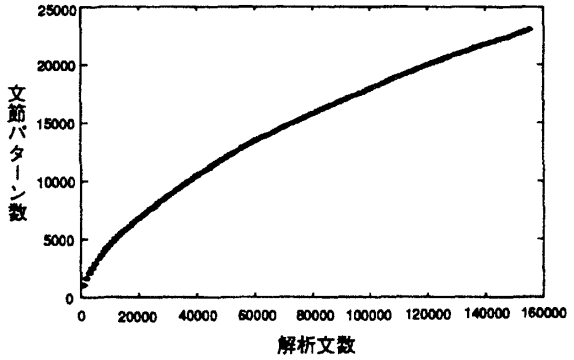

図 3: 解析文数と文節パターン数

次の例はコストデータの不備により失敗したものである．

『うっとうしくさせてきた』
解析結果：(うっとうし / く) (させ / てき / た)
正解　　：(うっとうし / く / させ / てき / た)

これは，文節コストデータベース作成時に，「A (形容詞) させてきた」が隣接部分との関係で解析があいまいとなり文節コストデータ抽出の対象になるものがなかったためである．

このように，文節コストデータ抽出の対象にならなかった部分を人手で選択して，正しいデータを取り込むことができれば，解析が間違いやすい部分に有効な質の高いコストデータが作成できる．人手が必要となるが，抽出する価値はあると思われる．また，他のコーパスからの抽出も行ないたい．

5　おわりに

単語と単語間にコストを付ける従来の方法に変えて，文節にコストを付け，大域的なコストを考慮した形態素解析システムを作成した．文節コストデータベースは，計算機により自動的に抽出した文節パターンの出現頻度をもとに作成した．

この方法により高精度の文節解析ができることが確かめられた．今後は文節コストデータの改良，パターン化規則の検討などをすすめていきたい．

参考文献

[1] 池戸，兵藤，奥村，栄留，池田：最長一致法に基づく 3 種のアルゴリズムを融合した形態素解析，言語処理学会第 2 回年次大会発表論文集 p.69-72, 1996.