文節単位のコストに基づく日本語文節解析システム

兵藤安昭　　池田尚志
岐阜大学工学部
{hyodo,ikeda}@ikd.info.gifu-u.ac.jp

1　はじめに

日本語文の係り受け解析では，文節を解析の単位として用いる．従って，形態素解析で単語単位に分割された文を文節単位にまとめあげる必要がある．文節は，1個の内容語と0個以上の機能語から構成されるという定義に従って，単語をまとめれば，文節の認識は容易に思える．しかし，例えば「解決策 として 浮上」の「し」のように本来の動詞としてではなく，まとまった機能語表現として使われる場合も多いため，細かく単語単位に分割してしまうと文節の認識はそれほど容易ではない．そこで我々は，係り受け解析を行う前段階としての文節解析システムを構築するために，まず初めに，文節を意識した長単位の機能語辞書の整備を行った．そして，分割された候補の中からより適切な文節を選択するために文節と文節間にコストを与える文節単位のコストを考慮したシステムの構築を行った．

2　システム概要

解析システムの流れは以下の通りである．

1. 辞書を参照して文節候補の作成
2. 各文節候補に文節カテゴリを付与
3. 文節にコストを付与
4. 文節間にコストを付与
5. コスト最小解を求める

解析用辞書，接続規則等はRDB上で管理し，容易に辞書編集ができるインタフェースを備えている．解析用の自立語辞書は，EDR日本語単語辞書 [1] をベースに平仮名以外の同一文字種からなる普通名詞・サ変名詞を削除したもの (約13万語) を用いている．本システムでは，漢字連続文字を名詞の可能性があるとして解析し，複合語処理は，文節が認定された後に行う．また，機能語辞書は，3節で述べるように独自に作成した．

パトリシア構造によるメモリ上の辞書を参照して，入力文中の各位置から始まる単語を切り出し，解析表に書き込む．そして，連接可能性をチェックし，文節内の末尾の単語が未然型ではないなどの規則により文節候補を作成する．次に，各文節に文節カテゴリを付与する．文節カテゴリは，構文的な観点から文節を分類したものであり，係り受け解析を行う際，係り可能な文節を求める情報として利用する．文節カテゴリは，現在，表1に示す75種類を設けている．例えば「体が用」とは，用言に係る体言文節を示し，再分類として助詞「が」が含まれていることを示す．そして，文節候補に対して，文節と文節間のコストを参照し，Viterbiアルゴリズムを用いてコスト最小解を求める．

表 1: 文節カテゴリ

体言文節：(15種類)
体の体, 体＊体, 体を用, 体に用, 体が用, 体と用, 体も用, 体は用, 体で用, 体か用, 体V用, 体＊用, 体＊並 体＊末 体＊＊
用言文節 (動詞, 形容詞, 形容動詞, 名詞述語文節): (11 × 4=44種類)
用＊体, 用Φ用, 用連用, 用終用, 用仮用, 用接用, 用名用, 用は用, 用名並, 用引用 用＊末
連体詞, 副詞, 接続詞文節：(3種類)
副＊体, 副＊用, 接＊用
引用文節：(3種類)
と＊用, と＊体, と＊末
その他：(10種類)
感動詞文節，用言命令型文節，括弧等
合計：75種類

3 文節解析のための機能語辞書

文節を意味的なまとまりに従って切り出すために，できるだけ長い単位で機能語を登録した．これは，以下のような理由からである．例えば「日本が変わるかもしれないと期待したが,」を「(日本が) / (変わるかも) / (しれないと) / (期待したが，)」と４文節に分割した場合，近接用言文節に係るという規則を用いて係り受け解析を行うと,「日本が」の係り先は「変わるかも」「しれないと」「期待したが,」の３つが考えられる．しかし「かもしれない」を１つの機能語として扱うことで，係り先のあいまいさを減らすことができる．また，機械翻訳など応用分野での意味の扱いに有利であることも予想される．

しかし，本動詞か機能語かが，あいまいな単語が文節に含まれていると，この方法では，本動詞と判断した場合には，文節を分割する必要がある．例えば「をはじめ/を始め」には，次のような出現例がある．

1. 「政府税調 をはじめ 大蔵省などは」
2. 「０．５ミリのボールベアリング を始め、驚くべき精緻な工作がなされている。」
3. 「国会で議論 をはじめ,」

3. のように本来の動詞として使われる場合には「を」と「はじめ」を分割しなければ正しい係り先を得ることができない．このように分割する可能性がある機能語については，辞書上にそのような情報を与えておいて，あいまい性解消の処理を行うことで，対処可能である．

3.1 機能語の整理，分類

上記のように機能語を長い単位で登録するという方針をとったが，機能語部分全体をすべて登録することは不可能である．そこで，機能語を基本的には終止型の単位で登録し，それらの間の接続規則を設定した．例えば,「(用言終止型) ＋にちがいないそうだ」は「にちがいない」＋「そうだ」のように登録する．このように長単位で扱うことによって，接続規則は，例えば「そうだ＋にちがいない」の生成を許さないようになっている．機能語は以下に示す５つに分類して登録した．

1. 基本機能語 (機能語後接)
   「ない，た」等
2. 体言後接機能語
   「が，を，に，に関して，について，をはじめ」等
3. 終止型後接機能語
   「が，けれども，かもしれない，はずがない」等
4. ている型機能語
   「ている, ていく, てから, てみる」等
5. 未然連用型後接機能語
   「ながら，そうだ，まい，ざるをえない，なくてはならない」等

3.2 複合機能語辞書記述

機能語には，例えば「ている，てはいる，てもいる」「かもしれない，かも知れない」のように，表記の違い，活用，助詞の付加による意味の添加などにより派生的な語が多数存在する．我々はスロット表現を用いて，このような複合的機能語の整理，収集を行い，辞書に登録した．

図１は，機能語「ざるをえない」とその派生語に対する辞書表現である．[@] をここではスロットと呼び，@1,@2 式がスロットに入り得る要素を示している．これにより「ざるをえない」のグループとして８個が表現されることになる。

```
ざるを @1 @2
          @1 え 得/え
          @2 ゆ ぬ ず まい
```

ざるをえ　　ざるを得
ざるをえぬ　　ざるを得ぬ
ざるをえず　　ざるを得ず
ざるをえまい　　ざるを得まい

図 1: 機能語スロット表現

4 文節単位のコスト最小法

形態素解析で一般的に用いられているコスト最小法では，個々の単語に与える単語コストと隣接する単語の接続に対する連接コストを用いて，総コストの少ない単語列を優先解として出力する (単語単位の方法)．これに対し，我々の手法では，個々の単語および単語間にコストを与えるのではなく，文節および隣接する文節間にコストを与えることとした (文節単位の手法)[図2].

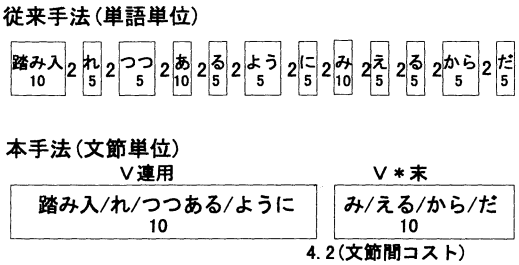

図 2: 文節間コスト

4.1 文節コスト

文節コストとしては，文節内のすべての機能語および機能語間のコストは考慮せず，内容語に関するコストのみを文節コストとして用いた．内容語コストは，単語の品詞 (品詞コスト) および単語の表記 (表記コスト) により決定する．表記コストは、EDR 単語辞書中の単語見出しには無い語で，そのカナ表記を辞書登録する場合に高いコストを与えている．

4.2 文節間コスト

文節カテゴリの bigram 確率値の対数 (の絶対値) を文節間コストとして設定した．この文節間コストデータを作成するのに，EDR 日本語コーパスを使用した [1]．EDR 日本語コーパスには約 208,000 文のテキストに対して，形態素情報，構文情報，意味情報が登録されている．そこで，EDR コーパス中に出現するテキストを文節コストのみを使って (文節間コストは使わずに) 解析し，EDR コーパス中の形態素情報と照合しながら正解判定を行う．こうして得られる正解解析文に対する文節カテゴリ列から文節カテゴリの bigram データを作成した．正解判定は，以下の通りである．

1. 文節区切り
   システムが出力した文節区切りから始まる単語が，EDR コーパスでは自立語 (接尾語, 語尾, 助詞, 助動詞, 以外) として登録されている．
2. 品詞選択
   システムが出力した文節の自立語品詞と EDR コーパスの品詞が同じである．

5 評価実験

EDR コーパス約 208,000 文のうち，100,000 文を文節間コストの学習用テキストに使用し，学習用テキストを Closed Text，残りの 100,000 文を Open Text としてシステムの評価を行った．結果を表 2 に示す．

表 2: 解析精度

	Closed Text		Open Text	
	文節切り	+品詞選択	文節切り	+品詞選択
なし	98.72%	95.74%	98.70%	96.14%
Bigram	98.73%	96.27%	98.72%	96.72%

文節間コストを考慮することで，品詞選択に関しては，0.5%程度，改善することができた．品詞選択の誤りとして，以下の例がある．

- 1 時間に 3 組の西洋人グループが入国を試み、いずれも 拒まれた。

EDR コーパスでは，この場合，「いずれも」は副詞となっているが，「名詞＋も」と解析する．これは「体も用→V＊末 (文節間コスト:1.2)」の文節間コストが「副＊用→V＊末 (文節間コスト:1.9)」のよりも小さいからである．

6 おわりに

本稿では，係り受け解析を行う前段階としての文節解析システについて述べた．まず初めに，文節を意識した長単位の機能語辞書の整備を行い，そして，分割された候補の中からより適切な文節を選択するために文節と文節間にコストを与える文節単位のコストを考慮したシステムの構築を行った．EDR 日本語コーパスに対する実験を行った結果，文節切りで，98.0%の正解率を得ることができた．

今後は，さらに機能語辞書の整備を行っていく予定である．

参考文献

[1] 日本電子化辞書研究所: EDR 電子化辞書仕様説明書, (1995)