文節解析のための長単位機能語辞書

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

1 はじめに

形態素解析システムでは，可能な単語候補を切り出し，接続可能なあるいは接続確率の高い単語列を選び出すという手順が一般的である．しかし，この方法では，完全な接続規則や接続確率を定めることが困難であるために，誤った機能語表現を生成してしまう可能性がある．

安武 [1] らは，日本語の連語データを人手で大規模に収集整理し，これらを同音異字の曖昧性解消に利用した仮名漢字変換実験を行っている．連語データ約 72,000 件のうち，「について，によって，なければならない」といった付属語性連語を 2,399 件登録している．また，山地 [2] らは連語を取り込むことで形態素解析の高精度化を図っている．我々は日本語文解析システム IBUKI を開発しているが [3]，そこでは，安武らの研究と同程度の比較的長い単位の機能語を基本機能語として辞書に登録し文節解析を行っている．

我々は，この考えをさらに押し進めて，特定の連語だけではなく，実際に現われる文節中の機能語列（以下，文節機能語列と呼ぶ）そのものを何らかの方法で，ほとんどすべて辞書に登録してしまうという方法の可能性について検討した．もし，すべての文節機能語列を辞書に登録することが可能ならば，機能語間の接続規則は不要になり解析処理は簡単になる．また，すべての文節機能語列が辞書に登録されることになるため，本当に辞書に登録すべきか否かを個々の機能語列毎に確認していくことができる．これによって解析の高精度化を期待できる．

我々の今回の分析結果では，すべての文節機能語列を数え上げることは不可能であるが，頻度上位 2,600 語の文節機能語列で総延べ語数の 99.0%を，頻度上位 27,000 語の文節機能語列で総延べ語数の 99.9%をカバーすることが分かった．これらの文節機能語列を従来の短単位の機能語辞書と合わせて登録することで，接続コスト等で高精度化を目指すのではない新たな文節解析システムを構築する可能性を得た．

2 機能語スロット辞書

我々が現在作成中の機能語辞書は比較的長い単位で機能語を登録しており，例えば「行くかもしれない」を１つの文節として扱い，文節内は「行+く+かもしれない」のように形態素に分割する．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る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，このような機能語を１つのグループとして扱い，機能語の整理・収集を行って辞書に登録した（以下，機能語スロット辞書と呼ぶ）．

図１は，グループ「ざるをえない」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．入りうる要素は，図に示すように [!] をつけることで，要素をさらに展開することもできる．これにより「ざるをえない」のグループとして 32(=2 × 16) 個が表現されることになる．

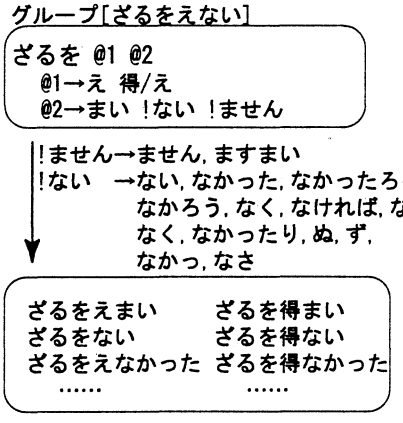

図 1: 機能語スロット辞書

現在の所，グループ数は 375，展開後の見出し語数は 13,882 である．機能語スロット辞書は以下に示すよ

うに，大きく 6 つに分類して登録している．

1. 体言後接機能語
   「が, を」等の格助詞に加えて「に関して, に際し, について」等の格助詞に相当する働きを持つ語を多数登録した (登録グループ数 : 150，グループを展開後の語数 : 2,164).
   例:「にかけて, にともなって, によると, における, をとおして, をかいして」
2. 用言未然形後接機能語
   「れる, られる, ない」等の用言未然形に後接する機能語を登録した.「ない」グループは，展開すると「ない, なかった, なかったろう, なかろう, なく, ず, ぬ, なければ, なくて, なかったり,...」等に展開される (登録グループ数 : 22，展開して得られる語数 : 576).
3. 用言連用形後接機能語
   「た, たい, ながら」等の用言連用形に後接する機能語を登録した (登録グループ数 : 34，展開して得られる語数 : 770).
   例:「うる, ます, なる, そうだ, する」
4. 用言連用形後接機能語 (ている型)
   「ている, てある」等，用言連用形に後接する「て」を含む連語表現を登録した (登録グループ数 : 45，展開して得られる語数 : 8,422).
   例:「ておく, てくる, てやる, てくれる, ていく, てもらう」
5. 用言終止連体形後接機能語
   「が, ので」等の接続助詞や「かもしれない」といった連語表現等，用言の終止連体形に後接する機能語を登録した (登録グループ数 : 91，展開して得られる語数 : 1,794).
   例:「どころではない, べきではない, とはかぎらない」
6. 用言終止連体形後接機能語 (形式名詞)
   「こと, ため」等，用言の連体形に後接する形式名詞を登録した (登録グループ数 : 25，展開して得られる語数 : 120).
7. その他
   用言仮定形や副詞に後接する機能語を登録した (登録グループ数 : 8，展開して得られる語数 : 36).

京都大学コーパス [4] を利用して，機能語スロット辞書を用いた文節区切りに対する評価実験を行った．京都大学コーパス (19,955 文) 1 文あたりの平均文節数が 9.64 に対して，我々のシステムでは平均文節数が 9.19 であった．京都大学コーパスでは,「〜に関して，〜について」等を「〜に/関して，〜に/ついて」と 2 つの文節としているため，我々のシステムの平均文節数が少なくなったことが考えられる．そこで，今回はシステムが出力した文節区切りから始まる単語が，京都大学コーパス上で自立語として登録されている場合を正解として判定した．その結果，正解率は 99.5%であった．

表 1: 頻度順位毎の文節機能語列統計

頻度の順位	延べ語数	カバー率	平均文字長
1 〜 100	23,801,259	91.18	1.28
101 〜 200	700,597	93.86	3.54
201 〜 300	341,385	95.17	3.51
301 〜 400	212,619	95.99	3.89
401 〜 500	142,317	96.53	3.72
501 〜 1000	351,057	97.88	4.43
1,001 〜 2,000	232,445	98.77	5.08
2,001 〜 5,000	173,939	99.43	5.82
5,001 〜 10,000	68,444	99.70	6.70
10,001 〜 20,000	41,148	99.85	7.47
20,001 〜 40,000	26,323	99.95	8.53
total(51,913)	26,103,446	100	1.55

$$\text{カバー率} = \frac{\text{累積延べ語数}}{\text{総延べ語数}}$$

$$\text{平均文字長} = \frac{\sum(\text{頻度} \times \text{文字長})}{\text{延べ語数}}$$

3 大規模データからの文節機能語列の抽出

3.1 新聞記事 4 年分に対する頻度統計

毎日新聞記事 94 年から 97 年の 4 年分 (4,073,889 文) に対して機能語スロット辞書を用いて文節解析し，文節機能語列の頻度を調べた．機能語を抽出する際に，例えば「学ばれている」「飲ませている」からは「ている」のみを文節機能語列とし,「れ」「せ」は抽出の対象にはしなかった．また，用言連用形に後接する「て，で」を含む機能語は「て」に統一し（例えば「でいく」は「ていく」とする），漢字表記はすべて平仮名表記に変換して統計をとった．

抽出した文節機能語列の総延べ語数は 26,103,446，異なり語数は 51,913，最大文字長は 24（「(融資し) ようということではなくなったのではないでしょうか」）であった．また，スロット辞書に登録した単語 13,882 エントリ中で，実際に出現した単語は 3,360 エントリであっ

た．頻度順位毎の文節機能語列の統計を表 1 に示す．

当初，新聞記事 4 年分という大量のデータを解析すれば，文節機能語列の異なり語数はある値に収束すると予測したが，頻度 1 の文節機能語列が異なり語数で 25,590 (49.3%) も存在し，この予測は成立しなかった．しかし，文節機能語列の総延べ語数に対する頻度上位 2,600 語の延べ語数（カバー率）で見れば 99.0%，上位 27,000 語の延べ語数で見れば，99.9%を占め，実際に出現する文節機能語列は「ほとんど有限」と考えることができる．

表 2: 文字長上位 3 位の文節機能語列

長さ	機能語列
24	ようということではなくなったのではないでしょうか
22	ということになってしまったのではございますが
22	てもらえなかったからというわけでもあるまいが
22	てしまったというようなものでもなかったはずだ
21	なければならないところであるにもかかわらず
21	なければならないということではないのだろう
21	なければいけないというようになっていますが
21	ではなくなってしまっているということである
21	できるようにしてもいいのではないでしょうか
21	てもらいたいということになるかもしれないが
21	ていることになっているのではないでしょうか
21	ていかなければならないのではないでしょうか
21	ていかなければいけないのではないでしょうか
21	ていかなくてはいけないということでしょうね

表 3: 分割した文節機能語列に対する統計

頻度数の順位	延べ語数	カバー率 (%)	平均文字長
1 ～ 100	25,747,295	93.29	1.31
101 ～ 200	841,372	96.34	3.17
201 ～ 300	367,087	97.67	3.33
301 ～ 400	196,079	98.38	3.94
401 ～ 500	117,563	98.81	3.81
501 ～ 1000	231,901	99.65	4.28
1,000 ～ 2,000	76,503	99.93	5.01
2,000 ～ 5,000	18,810	100.00	5.90
total(6,261)	27,597,871	100.00	1.46

3.2 文節機能語列の分割による頻度統計

3.1 節の結果から，実際に現れる文節機能語列をすべて数え上げることは難しいことが分かった．低頻度語は，表 1 からも分かるように文字列長が大きく，また，多くの場合，表 2 のように連体形機能語に「こと」「の」等の形式名詞や引用表現「という」を加えて新たな機能語が接続することで長い文節機能語列となる場合が多かった．そこで，次に，文節機能語列を意味的な切れ目と考えられる箇所で分割することを考えた．そうすれば，当初の予想が成立するかと期待し同様の統計をとることとした．具体的には終止連体形の箇所を意味的な切れ目とした．例えば「(融資し) ようということではなくなったのではないでしょうか」は「よう」「という」「ことではなくなった」「のではないでしょう」「か」と 5 つに分割する．頻度順位毎の統計を表 3 に示す．述べ語数は 27,597,871，異なり語数は 6,261 となったが，やはり，頻度 1 の分割した機能語列が 1624(25.9%) も存在した．しかし，カバー率では，上位 2,000 語で 99.93%とさらに高い「有限性」がみられた．

頻度 1 の文節機能語列中では，「こと, の, もの」等の形式名詞を含む機能語が多く存在していた．以下に例を示す．

- のにかんして
  山陽新幹線が新大阪―西明石間で壊滅的な打撃を受けた のに関して，…
- のばかりである
  雑誌は破天荒に楽しいのがなく，辛気くさい のばかりである．
- さいについては
  今後、陵墓を造る 際については「(葬儀までの間に) 時間的余裕がないのなら，事前に調査をするのも一つの手」と，事前の発掘調査も検討する考えを明らかにした．
- さいよりも
  また，被保険者である子供が死亡した 際よりも，契約者である親が死亡した場合の保障の方が手厚いのも大きな特徴だ．

これらに対して，「ことにかんして」は新聞記事 4 年間中で延べ 62 回，「ことばかりである」は 6 回，「さいに」は 1,883 回，「ときよりも」は 41 回出現していた．今回の実験では，形式名詞を含む連語表現も抽出したかったため，形式名詞を機能語として文節中の機能語列を抽出し数え上げることを試みたが，例えば形式名詞を自立語として文節を分割することで，文節機能語列の有限性を高めることができるかもしれない．しかし，このように形式名詞を機能語と考え，より大きな単位で文節を捉えることで，通常ではあまり使わない表現を指摘することができる可能性もあり，文書校正支援に活用できるのではないかと考えている．

4 おわりに

本論文では，我々が現在作成中の機能語スロット辞書について述べ，この辞書を用いて毎日新聞記事4年分を解析し文中に実際に現れる機能語列を収集し，その頻度統計について報告した．新聞記事4年分から抽出された文節機能語列は異なり語数で51,913で，その内，頻度1の文節機能語列が異なり語数で25,590（49.3%）も存在し，文節機能語列をすべて数え上げることは困難であることを確認した．しかし，文節機能語列の頻度上位2,600語で見れば，総延べ語数の99.0%を，上位27,000語で見れば99.9%を占め，実際に出現する文節機能語列は「ほとんど有限」であると考えて良いことが分かった．これらの文節機能語列を従来の短単位の機能語辞書と合わせて登録することで，接続コスト等で高精度化を目指すのではない新たな文節解析システムを構築する可能性を得た．今後は，これらの文節機能語列について，本当に辞書に登録すべきか否かを個別に確認して辞書を構築し，解析システムに適用したいと考えている．

参考文献

[1] 安武満佐子, 小山泰男, 吉村賢治, 首藤公昭: 日本語連語データとその応用, http://unicorn.tl.fukuoka-u.ac.jp/~yasutake/symposium.html, 「言語資源の共有と再利用」シンポジウム (1999)

[2] 山地治，黒橋禎夫，長尾 眞：連語登録による形態素解析システム JUMAN の精度向上, 言語処理学会 第2回年次大会, pp.73–76 (1996)

[3] 兵藤安昭, 池田尚志：文節単位のコストに基づく日本語文節解析システム, 言語処理学会 第5回年次大会, pp.502–504 (1999)

[4] 黒橋禎夫：構文情報付きテキストコーパスの作成と構文解析システムの改良, 言語処理学会 第5回年次大会 ワークショップ論文集, pp.57–62 (1999)