大規模コーパスからの未登録漢字列の抽出と辞書作成

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

1　はじめに

自然言語処理において未登録語の出現は避けられない。特に固有名詞や略語は未登録語である場合が多い。未登録語があるために解析を失敗することも多いので、未登録語を正しく抽出することは、形態素解析の精度を高めるための重要な処理の一つである。

また我々は自動点訳システム（IBUKI-TEN）を開発しているが、点訳では漢字に対して正しい読みを与える必要がある。未登録語に対しても正しい読みを与えなければならないし、略語であれば省略しない元の形の語も添えて置くべき場合もある。自動点訳後の後編集の際に未登録語を抽出し点検を促してくれれば効率的である。

本報告では、我々の文節解析システムIBUKIの解析結果を分析して、漢字未登録語を抽出する試みについて述べる。文中に存在する未登録語は、解析システムによってうまくそのまま未登録語として解析される場合と、分解されたり隣の語と結合してしまってそのままの形では解析されない場合とがある。たとえば「../福建/、/..」は「福建」をそのまま未登録語として解析しているが、「../西/武/の/..」では「西」と「武」をそれぞれ1文字の単語として解析してしまっている。また、「../買い/物/客/ら/で/..」では「買い物が辞書になかったために「/物/客/」がそれぞれ1文字の単語として解析されてしまっている。これらに関して「福建」「西武」等を未登録語候補として抽出する事が目的である。

実際に毎日新聞記事8,699,490文(9年分)を文節解析して、その解析結果を分析したところ、約17,000語を登録単語候補として抽出できた。この登録単語候補を手作業で点検した結果、その94％(約16000語)は未登録単語であり、残りの6％(約1000語)は非単語であった。なお、辞書としてはEDRの単語辞書を用いており、したがってEDRに登録されていない単語を抽出することになる。

2　大規模コーパスの形態素解析

IBUKIは文節の列を解析結果として出力する。文節は自立語部分と機能語部分からなる。辞書にない漢字列、カタカナ列、英数記号列はまずとりあえず‘未登録漢字列’、‘未登録カタカナ列’、‘未登録英数記号列’としていずれも名詞の扱いで解析され第1段階の文節解析がなされる。未登録漢字列、未登録カタカナ列は次の段階で複合語解析を受け単語に分割されて最終的な文節解析結果となる。

毎日新聞記事9年分（8,699,490文）をIBUKIで文節解析した。解析結果の統計を表1～4に示す。表1は未登録漢字列、未登録カタカナ列、未登録記号列に関する統計である。表2は未登録漢字列の文字列長に関する統計である。表3は複合語解析した結果についての統計である。表4は複合語解析した結果の漢字列の文字列長に関する統計である。

字種	述べ	異なり
全体	23857972	2341921
漢字	13158867	2073536
カタカナ	4417578	175327
英数記号	6281527	93058

表1: 新聞記事9年分中の未登録文字列

3　未登録漢字列からの登録単語候補の抽出

3.1　抽出

表4中の未登録漢字列がそのまますべて単語であるわけではない。また逆に表4中の未登録漢字列以外の漢字列の中に未登録の単語がないわけではない。解析の誤りのためにそのようなことがしばしば発生する。たとえば「.. 関越道..」→　「../関/越道/..」では登録

文字数	述べ	異なり
全体	13158867	2073536
1	61019	996
2	814031	44585
3	2548626	208320
4	5084076	564733
5	1902619	392042
6	1295013	332061
7	614248	193275
8	367508	125057
9以上	471727	212467

表 2: 未登録漢字列の文字列長

字種	述べ	異なり
漢字 (未登録)	307121	42239

表 3: 複合語解析後の漢字列

すべき未登録語「関越」が解析誤りによって未登録語として現れない。

今回は、漢字2文字および3文字の未登録単語を抽出することを目標として、まず表4中の長さ2または3の未登録漢字列（たとえば /熱女/ /阪和線/)、および表5の複合語解析の結果1文字ずつの2文字または3文字連続となった文字列（たとえば /邦/銀/ /刑/務/官/）を登録単語候補とした。

そのような登録単語候補は約87,000語あったが、以下の処理によって候補を絞り込んだ。

3.2 絞込み

まず、表4の長さ2または3の未登録漢字列の中で、その周囲の状況から単語である可能性が低いものを除外した。この単語らしさの条件は、次のいずれかを満たすパターンが解析結果9年分中に1度でもあることである。

1. その前後に同一文節内の単語がない
2. 前後に単語があっても機能語、未知カタカナ、未知ローマ字、記号、句読点である

また表5の複合語解析の結果1文字づつの2文字または3文字連続となった文字列の中で、1文字づつの分割で正解である可能性が高いものを除外した。この

文字数	述べ	異なり
全体	307121	42239
1	30624	1013
2	267637	38480
3	8239	2467
4	559	249
5	62	30
6以上	0	0

表 4: 複合語解析後の未登録漢字列の文字列長

文字数	述べ	異なり
2文字	385974	30804
3文字	122014	21103

表 5: 複合語解析で1字ずつに分割される漢字列

分割の正しさの条件は、次のいずれかを満たすものとする。

1. 「同..」「..氏」等のように1文字単語の前後にもよく現れる接辞を含んでいる
2. 「数詞」＋「数量後接語」という品詞構造である

最後に、出現数が5回未満の未登録漢字列を候補から除外した。ここで、表4と表5の単語には重複が有り得る (例えば/西武/ と /西/武/)。この条件での出現数とは抽出した単語の表4と表5での出現数を合計したものを指す。

3.3 最終的な登録単語候補

前述の処理の結果、17,076語の文字列が得られた。これが今回抽出した登録単語候補である。その出現数上位20語を表6に示す。

4 登録単語候補の検査

17,076語について人手で単語であるか否かの検査を行った。単語と認められるものには分類属性を付与し、略語に対しては点訳を考慮して元の読みが必要と思われるものにはそれも付与した。その結果を表7に示す。未登録単語であると推定したもののうち約94％(16093語) は単語であった。適合率としては高い値が得られ

語	出現数	文節例
西武	7869	/池袋/西武/、/
住専	6838	/住専/や/
近鉄	6667	/近鉄/バファローズ/の/
社民党	6412	/社/民/党/を/
米側	4745	/米/側/も/
社民	3661	/村山/富市/社民/党/党首/か/、/
長銀	3571	/長銀/新宿/支店長/の/
阪急	3221	/阪急/、/
計約	2743	/計/約/１７００億/円/の/
乃花	2724	/○/貴/乃花/・/魁皇/
三菱	2132	/同年/三菱/日本/重工業/
制球	2085	/制/球/ミス/が/
昨季	2063	/昨季/は/
野茂	2045	/イチロー/対/野茂/の/
郵貯	1776	/郵貯/の/
勧銀	1728	/勧銀/の/
死一	1612	/１/死/一/、/
米兵	1593	/米/兵/を/、/
信組	1591	/同/信組/守口/支店/が/
米韓	1540	/米韓/と/

表 6: 候補上位 20 語

たといえる。再現率についてはデータが大量であるので算出は困難である。表 7 の結果をみると、特に人名を代表とする固有名詞が大部分であった。

また略語 57 語について、省略しない元の形の語の読み情報を付加した。表 8 にその一部を示す。

5 おわりに

毎日新聞記事 9 年分を文節解析システム IBUKI で解析した結果から解析用辞書に未登録の 2 または 3 文字漢字列を抽出、その一部に対し単語であるか否かの検査を行い分類属性を付与した。この結果単語であったものを実際に解析用辞書に登録することによって特に固有名詞を含む文節について解析の正解率が上昇することが見込まれる。今後の課題として、今回文節解析の誤りで抽出できなかった未登録語や、4 文字以上の未登録語の抽出にも取り組みたい。これらは今回の結果が反映された辞書で解析した後に行うべきである。

品詞	数	例
人名 (姓)	2558	野茂　麻原
人名 (名)	1863	秀征　利休
人名 (その他)	2169	武双山　毛沢東
地名	417	雲仙　四川
地名 (海外)	384	重慶　広東
組織名	615	西武　阪急
年号	27	後漢　万暦
自然名	63	名瀑　白良浜
その他の固有名詞	2950	自自公　鳳鳳
普通名詞	952	制球　本戦
人称名詞	269	被葬者　父娘
時詞	244	戌年　朝型
その他	3580	馬連　茶髪

表 7: 品詞分類

語	読み	略語
勧銀	かんぎん	かんぎょうぎんこう
科技庁	かぎちょう	かがくぎじゅつちょう
農水	のうすい	のうりんすいさん
安企部	あんきぶ	あんぜんきかくぶ
電発	でんぱつ	でんげんかいはつ

表 8: 略語例

参考文献

[1] 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)