１文字未知語を核とする未知語候補の抽出

石原 吉晃，山田 佳裕，松本 忠博，池田 尚志
岐阜大学工学部

1　はじめに

コンピュータでテキストを解析するとき, テキスト中に現れる未知語（辞書に登録されていない語）を見出すことは重要な課題である. 未知語の抽出に関してはこれまでにも多くの研究がある [1][2]. それぞれ特徴があるが完璧な方法というものはない.

我々の研究室では文節構造を解析する ibukiC というシステムを開発している [3]. この ibukiC は, 辞書に登録されていないカタカナや漢字, アルファベットなどの連続部分は未知語文字列として切り出すが, そのほかに１文字の文字を出力する場合がある. このような「１文字未知語」が出現するのは, 多くの場合, 解析が乱れたために真の未知語が小さく分割されてしまった結果である. そこで本研究では, 大量のテキストの解析結果から, このような「１文字未知語」の周辺の文字列を切り出し，比較することで１文字未知語を核とする真の未知語の抽出を試みた.

本論文では, １文字未知語に関する詳細なデータ作成と, そのデータを元にした未知語候補の抽出法を提案し, この方法から得られた結果が有用であることを示す. 以下の節では, １文字未知語の具体例, データ作成の手順, そして１文字未知語を核とする未知語候補の抽出の具体的方法について述べ, 最後に本手法による抽出結果とその有用性について述べる.

2　１文字未知語

文節構造解析システム IbukiC で「やまたのおろち」という文字列を解析すると次のような結果となる.

> や（未知語/ひらがな）＋また（副詞）＋の（機能語）＋おろ（名詞/一般）＋ち（未知語/ひらがな）

本来「やまたのおろち」という一つの未知語として出力されることが理想的であるが, 誤解析により１文字にまで分割され, １文字未知語が出現する例である. このように真の未知語がひらがなで構成されている場合, ひらがなの１文字未知語が出現する可能性がある. 「しょうゆ」「わいろ」などの単語も同様で, 辞書に登録されていなければ１文字未知語が現れる.

また, 未知語が異なる文字種で構成されており, その要素の一つが１文字である場合にも, その単語は１文字未知語として解析される場合がある. 次は,「Ｙ２Ｋ」という名詞を文節構造解析システム IbukiC で解析した例である.

> Ｙ（未知語/アルファベット）＋２（未知語/数字）＋Ｋ（未知語/アルファベット）

この単語は「Ｙ２Ｋ」という未知語として抽出されることが望ましいが, 文字種が異なっているためそれぞれが１文字未知語として切り出されてしまう. 「貴ノ浪」「Ｇ８」などの単語も同様で、辞書に登録されていなければ１文字未知語に分割される.

今回の実験で使用した２０００年の毎日新聞記事において，このような１文字未知語がのべ 221,548 個出現している. その内訳は次の表の通りである.

表 1: １文字未知語数

ひらがな	カタカナ	アルファベット	漢字
87,538	15,214	36,275	80,192

3　１文字未知語データ作成

3.1　１文字未知語周辺文字列の切り出し

１文字未知語を手がかりとして未知語候補の抽出を行うためには, １文字未知語とその前後の文字列が必要となる. 本研究では文節構造解析システム IbukiC による解析結果の, １文字未知語を含む文節の前後１文節ずつを切り出して使用した. まず解析結果のなかから１文字未知語を検出する. その１文字未知語を含む文節を, 真の未知語抽出のための核文節とする. もし１文字未知語が隣接文節に含まれる場合は, その文節を含めて核文節とする. 核文節が確定した後, 核文節の前後の文節を切り出し, これを結合し１文字未知語を含む未知語候補文字列とする.

このテキストから抽出した未知語候補文字列を，１文字未知語部, 前文字列部, 後文字列部の三つの部分に分割し, 以下の処理を行う.

3.2　１文字未知語と前後文字列のソート

真の未知語を調べるためには，１文字未知語の前後にどのような文字列が接続しているかを調査する必要がある．

まず，抽出した文字列全てを１文字未知語部でソートする．次に１文字未知語部は固定したまま，１文字未知語に近い方から前文字列部をソートする．最後に，１文字未知語部・前文字列部を固定したまま，後文字列部をソートする．この処理により，１文字未知語の前後に似通った文字列が接続している未知語候補が存在した場合，それらは連続して出現することになる．

このソートの際，前文字列部の次に後文字列を処理した場合と，後文字列部の次に前文字列を処理する場合では結果が異なるが、これはどちらも有用な情報を持っている．そこで両方の結果を出力し，以下の処理に使用した．つまり以下で使用する１文字未知語の数は，テキストから抽出された数の倍となる．

3.3　一致文字列の切り出し

ソートした結果，１文字未知語の前後には，存在すれば同じ文字列が並ぶ．この前文字列部と後文字列部を隣接する文字列と比較し，一致する部分を切り出す．

ここは池　だ　研だ	→	だ
これは池　だ　研です	→	は池だ研
これは池　だ　研です	→	これは池だ研です
これは池　だ　研だろう	→	これは池だ研
ここは山　だ　部屋だ	→	だ

図 1: ソート後の文字列一致例

この切り出した部分が１文字未知語を核とする未知語候補である．一致する部分が存在したということは，少なくとも１度は同じ文字列がテキスト中に存在したということであり，真の未知語である可能性がある．この一致する部分が連続して出現した場合，その頻度が高いほど真の未知語である可能性が高いと考える．

この手法では，テキスト中の出現頻度が１の未知語については一切抽出できないことになるが，多頻度である未知語の重要度が高いと考え，テキスト中の出現頻度１の未知語に関しては本研究では取り扱わない．

4　未知語候補の抽出

本節では，前節で説明したデータを用いた，１文字未知語を核とする未知語候補抽出について述べる．

前節で抽出した１文字未知語の中で出現頻度の高いものが未知語候補と考えて良いが、それらの文字列を観察すると次のことがわかった．

1. 未知語の中に記号が含まれる可能性は低い
2. 未知語に助詞が付着していることがある
3. 高頻度未知語を含む文字列が低頻度未知語として少なからず出現する

この３点に着目し，さらに抽出処理を試みた．

4.1　未知語の中に記号が含まれる可能性は低い

未知語に限らず，記号を先頭，末尾，あるいは途中に含む単語は希である．そのため，記号が先頭及び末尾に在る場合はその記号を除去した．また途中に記号を含む場合は，その記号の部分で文字列を前後二つに分割し，１文字未知語を含む文字列を未知語候補とした．

次の例は「池だ」という人名の「だ」が１文字ひらがな未知語であるがめに，未知語候補として切り出された文字列を処理したものである．この処理により、文字列から有用な情報である「池だ」という文字列が未知語候補として抽出される．

（池だ・松本研究室）　→　池だ

この処理は繰り返し行い，データから記号は完全に消去するものとした．

4.2　未知語に助詞が付着している場合がある

１文字未知語を核とする未知語候補には，助詞が付着しているために異なる文字列と判断されているものが存在する．次の例は，本来は「幸せ」という未知語として抽出されることが望ましい未知語候補文字列の一部である．

が幸せ　の幸せ　幸せに　が幸せに　幸せ

これらは全て「幸せ（『せ』が１文字未知語）」という単語が真の未知語であるために抽出された文字列である．そこで，任意の助詞らしきひらがなが未知語候補の先頭あるいは末尾に存在している場合，そのひらがなを除去した．この助詞らしきひらがなとは「が」「の」「に」「を」などのひらがな１文字である．

しかし，ただ除去するだけでは，本来は助詞ではないひらがなまで除去されてしまう．そこで，未知語候補から助詞らしきひらがなを除いた文字列が，他の未

知語候補と重複する場合は, 除去したものが助詞である可能性が高いと判断した.

この手法では, 助詞らしきものが付着していても, 真の未知語がデータ中に存在していなければ, そのひらがなが助詞とは判定されないという問題点がある. また本来は助詞でなかったとしても, それを除いた文字列がデータ中に存在していれば, それが助詞だと判定されてしまう. しかし, 高頻度未知語候補に関しては多くの場合, 助詞らしきひらがなを除いた文字列がデータ中に存在しており、この手法は有用であると判断した.

最後に，この処理を行った後, 各文字列とその頻度も統合した.

が幸せ　の幸せ　幸せに　が幸せに　幸せ
→　幸せ：５

この頻度が多いほどデータ中に未知語候補が存在したことになり、その文字列が未知語である可能性が高いと考えられる.

4.3　高頻度未知語を含む文字列が低頻度未知語として少なからず出現する

前述の２手法である程度の未知語候補の抽出に成功した. しかし候補の数は多く, また切り出しの精度も高いとは言い難い. そこで, 高頻度未知語を含む文字列が低頻度未知語として多く出現している点に着目した.

例として, 阪神と阪神大震災を挙げる. これはどちらも抽出された単語であるが, 阪が１文字未知語であるためにどちらも未知語候補として抽出されている. つまり, 阪神大震災という未知語候補は, 阪神が未知語候補であるから出現しているのである. そこで, 高頻度未知語を含む体頻度未知語が存在した場合, その低頻度未知語は高頻度未知語と同じであるとした. 次に，頻度統合の例を挙げる.

阪神：１, ４４６　阪神大震災：４０８
→　阪神：１，８５４

この方法によって出力結果が整理され，より重要と考えられる未知語候補が高頻度となり，見易いデータとなった.

しかしこの手法では,「阪神大震災」という未知語候補は消えることとなる.「阪神大震災」は「阪神」という未知語のために出力されたとはいえ，これ自身が一つの未知語であるとも言える．４．３の処理は，抽出結果を見易くするという長所であるとともに,「阪神大震災」のような未知語を含む未知語を消去してしまうという短所を持つ. そのため，未知語を含む未知語の情報を得たい場合は，４．２までの処理結果に目を通す必要がある.

5　抽出結果及び未知語候補正解率

本節では毎日新聞記事一年分（2000 年:全 1,491,835 行）を対象として未知語候補抽出を行った結果と, 抽出された未知語候補の正解率について述べる.

5.1　未知語候補数

１文字未知語を核とする未知語候補の抽出は, ひらがな, カタカナ, アルファベット, 漢字という４種類の文字種毎に行った. 4.2 処理後と 4.3 処理後の抽出された未知語候補数を次の表に示す.

表 2: 抽出された未知語候補数

—	ひらがな	カタカナ	アルファベット	漢字
4.2 処理後	9,888	1,807	3,333	16,728
4.3 処理後	2,453	806	1,175	5,013

4.3 節で述べたように, 未知語を含む未知語など, 詳細な１文字未知語を核とする未知語候補を知りたい場合は 4.2 処理後の結果を見ることが有意である. しかし,4.3 処理を施す前後ではその未知語候補数が大きく異なる. そのため, 人手で単語を辞書に登録する際などは, 未知語候補を絞り込んだ 4.3 処理後のデータが有用であると考える.

5.2　抽出結果の正解率

4.3 処理により抽出された未知語候補のうち, 出現頻度の高かった単語１００を対象として評価を行った. 人が見て未知語候補と思われるものを正解として正解率を求めた. 正解率は次の表の通りである.

表 3: 抽出結果の正解率

ひらがな	カタカナ	アルファベット	漢字
32/100	49/100	50/100	90/100

5.3　抽出結果例

抽出結果の例として, それぞれの文字種で抽出された未知語候補の高頻度１０単語を次に示す.

表 4: ひらがなを核とした未知語候補

４．２処理後 (10,232 種)		４．３処理後 (2,468 種)	
未知語候補	出現頻度	未知語候補	出現頻度
など	3,444	など	5,314
だから	1,521	だから	1,760
幸せ	541	幸せ	1,016
少しずつ	390	から	838
だからこそ	239	いを	809
やむを得ない	236	いの	796
完ぺき	202	った	633
しょうゆ	197	りを	611
ふろ	195	りの	443
急きょ	194	って	440

表 5: カタカナを核とした未知語候補

４．２処理後 (13,410 種)		４．３処理後 (806 種)	
未知語候補	出現頻度	未知語候補	出現頻度
川崎フ	317	川崎フ	389
ガ大阪	308	ガ大阪	355
ヴ川崎	302	ヴ川崎	349
セ大阪	260	南ア	309
南ア	258	セ大阪	270
貴ノ浪	230	東レ	239
東レ	200	貴ノ浪	230
サ別	144	斗ヶ沢秀俊	200
琴ノ若	110	ヒ素	196
浜ノ嶋	107	サ別	149

表 6: アルファベットを核とした未知語候補

４．２処理後 (28,812 種)		４．３処理後 (1,175 種)	
未知語候補	出現頻度	未知語候補	出現頻度
W杯	1,045	W杯	1,488
W自前	767	W自前	767
W民新	547	W民新	547
ｉモード	351	Eメール	517
W民前	300	ｉモード	444
Eメール	250	G８	312
G７	223	G７	303
W自新	215	W民前	300
W社新	212	１Ｓ	271
Ｊリーグ	209	B組	253

表 7: 漢字を核とした未知語候補

４．２処理後 (70,942 種)		４．３処理後 (5,007 種)	
未知語候補	出現頻度	未知語候補	出現頻度
阪神	1,446	阪神	3,273
平壌	470	以外	1,804
阪神大震災	408	受けた	1,763
１億	386	百メートル	1,634
貴乃花	344	１億	1,426
受けた	301	以上	1,160
以外	283	平壌	830
選択肢	260	２億	762
二百メートル	254	０億	666
数百	252	数百	546

6　まとめ

文節構造解析システム IbukiC を用いて，１文字未知語を核とする未知語候補の抽出を行った．「しょうゆ」「わいろ」「ちほう」といったひらがな文字列であって１文字未知語に分割され誤解析となっていた単語や「Ｙ２Ｋ」「貴ノ浪」「Ｇ８」といった異なる文字種で構成されており, その要素の一つが１文字である場合の単語を抽出することに成功した．２節で述べた「やまたのおろち」のような単語も, 低頻度ながら出現を確認している．この抽出処理結果は, 解析用日本語辞書における未知語登録に十分役立つと考えている．

作成した未知語候補抽出プログラムは ibukiC の出力結果を入力すると，ibukiC が切り出した同じ文字種で構成されている未知語文字列，１文字未知語データ，４．２節までの処理を行った未知語候補，４．３節までの処理を行った未知語候補を全て出力することができる．

また, このプログラムは, 新聞記事１年分の文節構造解析システム IbukiC による解析結果を約３分で処理することができ, 実用性は高い．(CPU:Pentium4-2.4G MEM:512M で実験)

今後の課題としては, 抽出した単語の品詞推定の問題がある．本論文の手法では, 一致文字列を主に抽出したため, 抽出される品詞は名詞のみに止まらず,「やむを得ない」「いっぱい」のような副詞, 連体詞の類,「うーん」「あ～あ」といった感動詞なども抽出されている．また名詞にしても「斗ケ沢」や「貴ノ浪」といった人名,「千鳥ケ淵」「青木ケ原樹海」といった地名, また「ヴ川崎」「ガ大阪」といったスポーツチームの略称も抽出に成功している．

現在, これら抽出された単語を辞書に登録する際は, 人手で品詞の判定を行っている．今後, 品詞推定を同時に行うことで, より容易な辞書登録を支援するシステムの開発を検討していきたい．

参考文献

[1] 森, 長尾, ｎグラム統計によるコーパスからの未知語抽出, 情報処理学会論文誌,1998

[2] 池谷, 新納, 文字列が単語になる確率を用いた未知語抽出, 言語処理学会 第 6 回年次大会,2000

[3] 山田，高松，石原，水野，大口，佐藤，松本，池田, 日本語文解析システム ibukiC/S について, 言語処理学会 第 12 回年次大会,2006