新聞データからの「流行語」自動発見

―「コンピュータ流行語大賞」を目指して―

山川　侑吾[†]　　馬　青[†‡]
[†]龍谷大学
[‡]情報通信研究機構
qma@math.ryukoku.ac.jp

１　はじめに

「流行（語）」という言葉の持つ意味について、国語辞典で調べてみると、①一時的に急に世間にひろがりふえること、②移り変わること、とある。年一回、世間一般で用いられる「流行語」を決定する「ユーキャン流行語大賞」のホームページには“１年の間に発生したさまざまな「ことば」のなかで、軽妙に世相を衝いた表現とニュアンスをもって、広く大衆の目・口・耳をにぎわせた新語・流行語”と定義されている**[1]**。つまり、世間で比較的短期間に幅広く使用された言葉、ということができる。

さて、毎年１２月上旬に発表する「ユーキャン流行語大賞」を先取りし、毎年の「流行語」をコンピュータで予測することは可能であろうか。本研究はこのような疑問に答えるべく「コンピュータ流行語大賞」を作り出すことを目指すものである。

「流行語」を予測するためには、テレビ・雑誌・新聞・インターネットなどあらゆるメディア上の言語データを調査対象とし、その中から流行語になりうる言語特徴を用いて流行語を抽出（自動発見）する必要がある。本稿では、それを目指す第一歩として、各種メディアの中から最も文字情報が多く、かつ日常的に用いられる情報収集源である新聞のデータに焦点を絞り、そこから毎年の世相を端的に反映する「流行語」を抽出する試みを紹介する。

２　基本的な考え方

新聞データを形態素解析すると、辞書に登録されていない単語は「未知語」として出力される。一般には「未知語＝情報量なし」と考えがちであるが、未知語は新語、人名・地名・組織名である場合や、表記のゆれに起因する語句の一部である場合が多い。このような言葉や語句が短期間に高頻度で出現した場合、それらは「流行語」、または世相を端的に反映する言葉になる可能性が高いと考えられる。したがって、形態素解析の結果として望ましくないと思われる「未知語」を逆手に取り、それを手がかりに「流行語」を自動抽出できるのではないかと考える。

そこで、本研究では「流行語」の自動検出を目指し、新聞データの中からその「未知語」に絞って出現頻度を調べることにした。具体的には、単体では意味を成さないが、頻度の高そうな語句の一部分であるような未知語については、そのパターン別に異なる４種類の処理を施し、意味のある単語に変換し、出現頻度を調べることにした。

３　各種の未知語パターン

本研究では茶筌**[2]**を用いて新聞データの中の未知語とされた単語について調べてみると、それらは様々な種類に分類することができる。本節では、各種の処理を施す対象となった４種類の未知語について述べる。

３．１　人名漢字型

茶筌に登録されている漢字で構成されている人名の場合、品詞名は「名詞-固有名詞-人名」と出力される。しかし、あまり汎用的に使用されない漢字や、名前を構成する漢字の組み合わせが特殊な場合、各々が独立した

単語として解釈される。例えば、「裕也」と「民也」は同じ人名であるが、「裕也」は人名として登録されているために名詞として解釈されているが、「民也」は登録されていないので名詞「民」と未知語「也」として結果が出力される。

３．２　ＵＲＬ（アドレス）型

ウェブサイトのＵＲＬや、メールアドレスをそのまま茶筌に入力すると、記号は一つ一つが独立して「記号-一般」と解釈され、数字部分は「名詞-数」と解釈される（新聞では記号などはすべて全角で表示されている）。しかし、英文字部分は（記号や数字で区切られ）一括して「未知語」として解釈される。

例えば「ｈｔｔｐ：／／ｈｏｇｅ…」とあれば、茶筌では「ｈｔｔｐ」「：」「／」「／」「ｈｏｇｅ」と分けられ、解釈される。

３．３　漢字読み型

新聞データには、漢字なのに括弧書きで読み仮名が振ってある箇所がある。その際、「瞬（またた）く」のように読み仮名と送り仮名の間に（）が入っているためその箇所で分断されて、（）内の読み仮名は独立した平仮名の文字列として解釈される。茶筌では（）や各種記号はその用途如何に関わらず、記号として独立して解釈するため、上記の例では「またた」と「く」に分けられて解釈される。

３．４　「・」含みネーム型

「ハリー・ポッター」など、欧米の人名や固有名詞、およびカタカナ語の中には、単語と単語の間に「・」を含むものがある。その場合も、３．３の（）と同様に茶筌では「・」の前後で分断され、独立した単語として扱われる。

４　各種の処理

４．１　人名漢字型の処理

人名は、人名以外にも新聞で多岐にわたり使用されている漢字を処理の対象とするために、一般的な（四字）熟語や汎用的な固有名詞などと正確な区別をすることは非常に難しいが、未知語扱いされる人名は、少なくとも

１．漢字のみで構成されている
２．単体では「未知語」と判断される漢字を含んでいる

というような特徴がある。

このようなタイプの未知語に対しては次のような処理を施す。まず、データを読み込んだ際にそれが漢字であった場合、その漢字及び品詞名をリストに格納しておき、それ以降は入力が漢字であればその漢字と、漢字の品詞名をリストの最後尾に格納していく。そして、読み込んだデータが漢字でなくなったタイミングで、リストに格納してある品詞名を調べる。その中に一つでも「未知語」と判断された漢字が含まれていれば、そのリストに格納されている漢字群からなる単語を「人名」と判断して漢字群を結合させる。そして出来上がった漢字文字列を一つの未知語（＝「流行語」の頻度調査対象）として、出現頻度を調べる。

４．２　ＵＲＬ（アドレス）型の処理

ウェブサイトのＵＲＬやメールアドレスには特定の種類の文字列しか使われておらず、その配列にも以下のような規則性が見られる。

１．ＵＲＬ（アドレス）はアルファベット（小・大文字）または数字のいずれかから開始する。
２．ＵＲＬを構成する記号以外の文字は、アルファベット（小・大文字）と数字である。
３．記号は「：」「／」「．」「―」「@」である。
４．ＵＲＬ部分は文字列「ｊｐ」「ｈｔｍｌ」、または記号「／」で終了する。
５．「アルファベット（小・大文字）**or** 数字」「：**or**／ **or** ． **or** ― **or** @」の各グループは交互に出現する。
６．５の例外として、「：」の直後は「／」が２回連続で出現する。

すなわち、一般的に見られる「ｈｔｔｐ：／／www．ｈｏｇｅｈｏｇｅ…」や「ｈｏｇｅ@ｒｙｕｋｏｋｕ．ａｃ…」の形を想定して処理を行う。

このようなタイプの未知語に対しては次のような処理を施す。まず、データを読み込んだ際にそれが英文字もしくは数字であった場合、その単語をリストに格納しておく。その直後の単語が、記号「:」「／」「．」「ー」「@」のいずれかの場合、その記号をリストの最後尾に格納する。そしてそれらの直後、つまり［英文字 **or** 数字］→［記号］→［英文字 **or** 数字］の順になっている間、記号を次々とリストの最後尾に格納する。ただし、６に該当すればそちらの処理を優先させる。また、「／」はＵＲＬ記述の終わりの部分である可能性もある。そのため、「／」が出てきた場合はその次の単語が英文字か数字かによって、処理を続行させるか否かを判定する。処理を終了させる場合はそれまでにリストに格納した単語と記号を結合し、新しい単語を得る。一方、「ｊｐ」や「ｈｔｍｌ」といった単語が読み込まれた場合、それぞれメールアドレス、サイトのＵＲＬの終わりの部分を指すため、無条件で処理を終了し、単語の結合処理を行う。

４．３　漢字読み型の処理

このタイプの未知語は

１．読み仮名の直前の文字は「(」である
２．() 内は全て平仮名で構成されている
３．読み仮名の最後の文字の次は「)」である

というような規則性がある。

このようなタイプの未知語に対しては次のような処理を施す。データを読み込んだ際に、記号「(」が出ていればこの型の判定処理を始める。記号「(」以降、次に単語「)」が出てくるまで、全ての単語と品詞名を順に、リストに格納する。その際、途中で平仮名以外の単語があれば処理を中断して、それまでの（リストに格納されている）単語と品詞名に対し、通常の頻度調査処理を行う。一方、「(」以降、次に単語「)」が出てくるまで全て平仮名であった場合、リストに格納されている単語と品詞名を棄却し、処理を終了させる。

４．４　「・」含みネーム型の処理

「・」は、主に外国人の人名や地名などの固有名詞に用いられるが、「衆議院・参議院」のように、前後で並列または従属関係がある場合に用いられる事もある。本研究では外国人の人名や地名に焦点を絞る。このタイプの未知語は

１．名詞及び「・」のみで構成されている
２．一連の語句の最初と最後がカタカナである
３．単語と「・」が交互に現れる

というような規則性が見られる。

このようなタイプに対しては次のような処理を施す。データを読み込んだ際に、カタカナの名詞が出てきた場合、その単語をリストに格納する。その次の単語が記号「・」であった場合、リストの最後尾に格納する。このタイプの場合も４．３と同様、「カタカナ名詞」→「・」→「カタカナ名詞」の順になっている間、単語(記号)を次々とリストの最後尾に格納する。この処理を続けていき、リストに「・」が一つ以上格納されている状態でカタカナ名詞以外の名詞が出てきた場合、または「・」以外の記号が出てきた場合にはリストの単語を結合させ、結合処理を終了させる。

５　計算機実験

５．１　実験データ

本研究の実験には毎日新聞の新聞データ５年分（**1999** 年～**2003** 年分）を使用した。

５．２　実験結果

表１には **2003** 年の新聞データの処理結果を示す。左側の列に示している単語は上記各種処理を施した場合の結果であり、右側の列に示している単語は上記処理をしない場合

の結果である。単語の順位はそれぞれ総出現頻度（TF）、その単語を含む総記事数(DF)、そして上記を総合的に測る指数（TF+2×DF）を用いて示した。

この表より、上記各種の処理を施すことによってこの１年の「流行語」として「ＳＡＲＳ」、「ＮＧＯ」、「アルガイダ」、「ＩＴ」、「ラムズフェルド」を抽出することができたことが分かる。一方、上記処理を経ずに未知語に対する頻度調査を行ったところ、上位５個の高頻度単語は「ＳＡＲＳ」、「ｊｐ」、「ｃｏ」、「ｍａｉｎｉｃｈｉ」、「ｈｔｔｐ」であった。

また、他の４年分のデータの実験においても、上位５位までに例えば「ペイオフ」（1999年）、「ＩＴ」（2000年）、「タリバン、ビンラディン」（2001年）、「BSE」（2002年）（７位に「サッカー・ワールドカップ」）が入っており、2003年と同様、各年の世相を端的に反映している興味深い「流行語」が得られた。

５．３　考察

例えば1999年のデータからの「流行語」抽出においては「巨」という漢字が５位に入った。これは【巨】４－２【神】のように、スポーツ面の結果表示に用いる、チーム名の略称であった。このことから「特定の漢字一文字（球団名、学校名など）が一文字単位で出てきたら、無条件に排除する」という処理のプログラムを追加したほうが、下位の誤差が生じるといった弊害はあるかもしれないが、上位の検出精度としては向上できると思われる。

６　おわりに

本稿では、新語や表記のゆれに起因する「未知語」に焦点を当て各種の処理を施すことによって新聞データから「流行語」を自動抽出する手法を提案した。計算機実験の結果、興味深い「流行語」が検出できた。

今後、調査対象をよりタイムリーでより豊富な情報源であるウェブサイトなどに、抽出の対象を話し言葉も含めた語句に拡張するとともに、抽出の手がかりを多様化させ、「コンピュータ流行語大賞」の開発を目指す。

表１　２００３年の「流行語」抽出結果

	単語（処理あり）	TF	単語（処理なし）	TF
1	ＳＡＲＳ	3764	ＳＡＲＳ	3766
2	ＮＧＯ	1213	ｊｐ	2473
3	アルカイダ	842	ｃｏ	2003
4	ＩＴ	793	ｍａｉｎ	1844
5	ラムズフェルド	729	ｉｃｈｉ	
			ｈｔｔｐ	1336
	単語（処理あり）	**DF**	単語（処理なし）	**DF**
1	ＳＡＲＳ	3088	ＳＡＲＳ	3089
2	ＮＧＯ	1096	ｊｐ	2443
3	ＩＴ	701	ｃｏ	1996
4	ラムズフェルド	701	ｍａｉｎ	1838
			ｉｃｈｉ	
5	アルカイダ	688	ｈｔｔｐ	1314
	単語（処理あり）	**TF+2DF**	単語（処理なし）	**TF+2DF**
1	ＳＡＲＳ	9940	ＳＡＲＳ	10944
2	ＮＧＯ	3405	ｊｐ	7359
3	アルカイダ	2218	ｃｏ	5995
4	ＩＴ	2195	ｍａｉｎ	5520
5	ラムズフェルド	2131	ｉｃｈｉ	
			ｈｔｔｐ	3964

謝辞

本稿作成にあたって有益な助言をくださった内元清貴氏に感謝します。

参考文献

[1] 自由国民社：ユーキャン流行語大賞・全受賞記録（http://www.jiyu.co.jp/singo/）

[2] 松本裕治他：形態素解析システム『茶筌』version2.3.0使用説明書、2003