放送ニュースの基本語彙の抽出

美野　秀弥　　熊野　正　　田中　英輝

NHK 放送技術研究所　〒157-8510　東京都世田谷区砧 1−10−11

E-mail:　{mino.h-gq , kumano.t-eq , tanaka-h.ja}@nhk.or.jp

1. はじめに

著者らは、ニュースを国内の外国人住民のためにやさしく書き換えてサービスすることを検討している[1]。本稿では、このような目的に有用な「ニュースの基本語彙」の選定方法に関する研究を報告する。

ニュースをやさしくする手段には、主に単語をやさしくすることと構文をやさしくすることがあるが、その中でも単語をやさしくすることは書き換え作業者にとって大きな負担となっている。そこで著者らは、ニュースに頻出している単語をリスト化して、その意味を外国人住民にあらかじめ提示することができれば、これらをやさしくする必要がなくなり、書き換え作業の負担を軽減できると考えた。

次節で詳しく述べるが、このリストのことを本稿では「ニュースの基本語彙」と呼ぶ。著者らはこれを、効率よく、またできるだけ客観的に作るため、候補となる語彙を統計的にコーパスから抽出した後、専門家が精査することを考えている。本稿では、このために適した統計手法を実験的に検討した。

2. 基本語彙の考え方

ある分野にとっての基本語彙は、田中[2]によると、頻度が大きく、使用領域が広く、基本的な意味を持つものである。著者らはこの考えを参考にして、ニュース中の頻度が大きく、領域が広い語彙を「ニュースの基本語彙」とした。ただし、ここでの頻度や領域は人間の内省によるものと考えている。頻度や領域は数学的に定義ができるので、統計指標だけで適切な基本語彙リストを作ることができる。しかし田中[2]も指摘しているように、調査対象のコーパスによって偏りが生じる可能性があり、直感に反する語彙が選択されることがある。特にニュースを対象にした場合、一過性の事件や事故に影響された語彙が選ばれやすい。このような語彙は、その事件、事故以後に出現する可能性が低く基本語彙として適切ではないため、専門家が一語一語確認して取り除く必要がある。

もし、人間の内省による頻度や領域を反映した統計指標で候補語を抽出することができれば、取り除くべき語彙が少なくなり、効率よくニュースの基本語彙を作ることできるだろう。

すなわち、著者らの求める統計指標は、人間の内省に近い「頻度」や「領域」を反映したものである。このような指標を求める研究の第一歩として、本稿では基本的な統計指標を対象にその性質を調査した。

3. 関連研究

既存の基本語彙のリストに「日本語教育のための基本語彙調査」[3]や、「日本語能力試験出題基準」[4]などがある。「日本語教育のための基本語彙調査」[3]は、外国人の日本語学習者がはじめに学習すべき語を「基本語二千」「基本語六千」としてリスト化したものである。また、「日本語能力試験出題基準」[4]は、日本語能力試験のレベルごとの語彙をリスト化したものである。これらは主に人手で作られたものであり、自動的に抽出する手法についての言及はない。また、ニュースに頻出しない単語も数多く含まれており、ニュースの基本語彙としては必ずしも十分でない可能性がある。

統計指標を用いて基本語彙を抽出することを検討した研究に、内山ら[5]の研究がある。内山らは、初級、中級の英語学習者が特定分野の英語を効率的に学習するための語彙リストを作成するために、様々な統計指標を用いて語彙の抽出を行い、それらの指標が持つ特徴を分析している。しかし、内山らは特定の分野で特有に用いられる特徴単語を抽出することに主眼を置いている。そのため、一般的に用いられている単語も対象にしている本稿とは目的が異なる。

4. 基本語彙抽出のための統計指標

以下に、本稿で用いた基本的な統計指標を示す。

(i) 単純頻度と文書頻度

単語 w の出現回数(以下、単純頻度)と、単語 w を含む記事の数(以下、文書頻度)を「頻度」の指標として用いた。

(ii) KL 情報量

特定のジャンルと時期に偏らないで出現する単語の相対頻度分布は、全単語の相対頻度分布と同じになると考えてよい。そこで、ある単語と全単語の相対頻度分布の差を表す KL 情報量を、領域の広さを

表す指標として使うことを考えた。式(1)の KL 情報量は、全単語のジャンル g と時期 t ごとの相対頻度 $F(t,g)$と、ある単語 w のジャンル g と時期 t ごとの相対頻度 $G(w,t,g)$との差を表したものである。分布の差が小さければ 0 に近い値を取る。尚、$G(w,t,g)$ の分子 $c(w,t,g)$に 1 を足すことで $G(w,t,g)$が 0 にならないようにした。

$$Kl(w) = \sum_{t,g} G(w,t,g) \bullet \log\left(\frac{G(w,t,g)}{F(t,g)}\right) \qquad (1)$$

$$G(w,t,g) = \frac{c(w,t,g)+1}{\sum_{t,g} c(w,t,g)} \qquad F(t,g) = \frac{\sum_{w} c(w,t,g)}{\sum_{w,t,g} c(w,t,g)}$$

$c(w,t,g)$: ジャンル g と時期 t のニュース中の単語 w の単純頻度

(iii) 相乗平均

特定のジャンルと時期に偏らないで幅広くニュースをカバーしているかどうかは、単語 w のジャンル g と時期 t ごとの単純頻度 $c(w,t,g)$の相乗平均で測定することができる(式(2))。この指標は、領域の広さだけでなく、頻度の大きさも表している。

$$Ave(w) = \left(\prod_{t,g} c(w,t,g)\right)^{\frac{1}{TG}} \qquad (2)$$

T:時期 t の総数　G:ジャンル g の総数

5. 評価実験

5.1.実験概要

2000 年から 2010 年までの NHK ニュースを用いて、4 節で示した基本的指標間を比較、評価した。

これらの指標間の差は必ずしも大きくないため、基本語彙の抽出に用いられることの多い単純頻度で作成した語彙リスト(以下、高頻度語リスト)をベースラインにして、基本的指標(以下、比較指標)で作成した語彙リストと比較した。また、評価は筆者が主観的におこなった。

詳細な手順は次の通りである。

1) 単純頻度を用いて高頻度語リストを作成する。
2) 比較指標を用いて語彙リストを作成する。
3) 1),2)の語彙リストから、上位 500 語と 6000 語を抽出してリスト化する。これは、リスト内の単語数の違いによる影響を分析するためである。
4) 高頻度語リスト上位 500 語のみに出現した単語と、比較指標の語彙リスト上位 500 語のみに出現した単語から 1 語ずつ無作為に抽出した評価用ペアを 50 組作る。上位 6000 語のリストについても同様に評価用ペア 50 組を作る。
5) 評価用ペアの各単語を「主観的頻度の大きさ」と「主観的領域の広さ」の２つの基準で比較し、良いと評価した単語の指標を選ぶ。差がない場合は「比較不可能」とする。
6) 5)の結果を表 1 に当てはめて評価用ペア 50 組の総合評価を決める。
7) 6)で総合評価の数が多かった指標を「ニュースの基本語彙」の抽出に適した指標とする。

頻度の大きさ	領域の広さ	総合評価
単純頻度	単純頻度	単純頻度
単純頻度	比較不可能	単純頻度
比較不可能	単純頻度	単純頻度
単純頻度	比較指標	判定不可能
比較不可能	比較不可能	判定不可能
比較指標	単純頻度	判定不可能
比較指標	比較不可能	比較指標
比較不可能	比較指標	比較指標
比較指標	比較指標	比較指標

表 1：総合評価のための規則

以下、5.2.1.節では既存の基本語彙との比較と評価を行い、その特徴を分析した。5.2.2.節、5.2.3.節、5.2.4.節では 4 節で提案した比較指標の評価を行った。そして 5.3.節ではニュースの基本語彙の抽出に適した指標を考察した。

尚、本実験では、基本語彙として自明な助詞、助動詞、数などは対象から除外した。

5.2.指標毎の比較

5.2.1.既存の基本語彙表との比較

高頻度語リストと既存の基本語彙表を比較した。既存の基本語彙表には、「日本語教育のための基本語彙調査」[3]に掲載されている 2030 語と 6060 語の語彙リストを用いた(以下、教育基本語彙)。単語数を合わせるため、高頻度語リストの単語数も上位 2030 語と 6060 語とした。高頻度語リストと教育基本語彙の間で重複しなかった単語は上位 2030 語では 1316 語、6060 語では 3115 語あった。これらを用いて作成した評価用ペア 50 組を比較、評価した(表 2)。

表 2 は「高頻度語(リスト)」、「教育基本語彙」、「判定不可能」の 3 列で構成されている。そして、各列の中の「頻度(主観的頻度の大きさ)」と「領域(主観的領域の広さ)」の列は評価用ペア 50 組のうち、それぞれの基準で良いと評価された数を示しており、「総合」の列には「頻度」「領域」の評価結果を表 1 に当てはめて総合的に良いと評価された評価用ペアの数を示す。また、「判定不可能」の列には、「頻度」「領域」で「比較不可能」、「総合」で「判定不可能」となった数を示す。

表 2 の総合評価を比較した結果、2030 語と 6060 語ともに高頻度語リストの評価が良かった。これは、評価例(表 3)の「アイロン」、「洗濯機」のような、ニュースにはほとんど出現しないと評価された単語が教育基

本語彙の側に多く含まれていたためである。また、各語彙リストを数値データから評価するために、各語彙リストの単語 2030 語がニュースをカバーする割合を調べた結果、高頻度語リストのカバー率が 63.4%だったのに対し、教育基本語彙のカバー率は 24.8%と低かった。そこで、教育基本語彙の単語の出現数の分布を調べたところ、「梅干し」、「いらっしゃいませ」などニュースにほとんど出現しない単語も含まれていた。

以上から、教育基本語彙はニュースの基本語彙として必ずしも十分ではないと考えられる。

	高頻度語			教育基本語彙			判定不可能		
	頻度	領域	総合	頻度	領域	総合	頻度	領域	総合
2030	39	25	39	0	3	1	11	12	10
6060	22	20	22	1	12	9	27	18	19

表 2：高頻度語リスト-教育基本語彙の比較結果

高頻度語	外交	当初	事例
教育基本語彙	アイロン	洗濯機	車
頻度/領域	高頻度語/高頻度語	高頻度語/高頻度語	教育基本語彙/比較不可
総合	高頻度語	高頻度語	教育基本語彙

表 3：高頻度語リスト-教育基本語彙の評価例

5.2.2. 文書頻度との比較

単純頻度と文書頻度を比較した。各指標で作成した語彙リスト間で重複しなかった単語は、上位 500 語では 53 語、上位 6000 語では 306 語あった。これらを用いて作成した評価用ペア 50 組を比較、評価した (表 4)。

その結果、上位 500 語と 6000 語ともに文書頻度の評価が良かった。評価用ペアをみると、表 5 の評価例のように、文書頻度では「良い」、「残る」など用言が多かった。そこで、重複しなかった単語の品詞分布を調べたところ、高頻度語リストのみに出現した単語の品詞はほとんどが名詞だったのに対し、文書頻度の語彙リストのみに出現した単語の品詞は上位 500 語で 41.5%、6000 語で 29.4%が用言であった。用言は名詞に比べて様々な場面で使われることから、主観評価では用言が基本語彙と評価される傾向にあった。これが、文書頻度の評価が高かった一因と考えられる。

	単純頻度			文書頻度			判定不可能		
	頻度	領域	総合	頻度	領域	総合	頻度	領域	総合
500	4	4	4	29	39	38	17	7	8
6000	10	5	5	32	40	36	8	5	9

表 4：単純頻度-文書頻度の比較結果

単純頻度	爆発	議会	警部
文書頻度	良い	残る	阻む
頻度/領域	文書頻度/文書頻度	文書頻度/文書頻度	単純頻度/文書頻度
総合	文書頻度	文書頻度	判定不可

表 5：単純頻度-文書頻度の評価例

5.2.3.重み付き KL 情報量との比較

5.2.2.と同様の手法で、単純頻度と KL 情報量(式(1))を比較、評価した。尚、式(1)にあるジャンルには、「社会、暮らし・文化、科学・医療、政治、経済、国際、スポーツ」の 7 つを使った。また、時期の単位は月とした。

評価の結果、全ての組において、単純頻度が適切と評価された。KL 情報量で作成した語彙リストのみに出現した単語の単純頻度を調べたところ、そのほとんどが 10 未満であった。

そこで、低頻度語の影響を緩和するために、KL 情報量と単純頻度を掛け合わせた重み付き KL 情報量(式(3))を作成し、これを新たな指標として評価を行った(表 6)。その結果、上位 500 語では重み付き KL 情報量の評価が良かった。しかし、上位 6000 語では低頻度語の影響により単純頻度が適切と評価された。尚、各指標で作成した語彙リスト間で重複しなかった単語は上位 500 語では 102 語、6000 語では 781 語あった。

ここで、上位 500 語から作成した評価用ペアをみると、表 7 の評価例のように、高頻度語リストのみに出現した単語は「首脳」が政治、「投資」が社会など、ジャンルに偏りがあり領域が狭いと評価されるものが多かったが、重み付き KL 情報量で作成した語彙リストのみに出現した単語には「直接」、「テレビ」などジャンルや時期を推測できず領域が広いと評価されたものが多かった。このことから、重み付き KL 情報量が「主観的領域の広さ」を反映した指標であることを確認できた。しかし、重み付き KL 情報量で得られた上位 6000 語には単純頻度が 10 未満の単語が含まれており、低頻度語の影響を抑えられなかった。さらに低頻度語の影響を緩和するためには、低頻度語をフィルタリングするなどが必要となる。

$$Kl_T(w) = Kl(w) \bullet \sum_{t,g} c(w,t,g) \qquad (3)$$

	単純頻度			重み付き KL			判定不可能		
	頻度	領域	総合	頻度	領域	総合	頻度	領域	総合
500	4	6	3	40	44	41	6	0	6
6000	50	49	50	0	0	0	0	1	0

表 6：単純頻度-重み付き KL 情報量の比較結果

単純頻度	投資	首脳	表明
重み付き KL	直接	基づく	テレビ
頻度/領域	重み付き KL/重み付き KL	重み付き KL/重み付き KL	比較不可/単純頻度
総合	重み付き KL	重み付き KL	単純頻度

表 7：単純頻度−重み付き KL 情報量の評価例

5.2.4.相乗平均との比較

単純頻度と相乗平均(式(2))を比較、評価した(表 8)。その結果、上位 500 語と 6000 語ともに相乗平均の評価が良かった。尚、各指標で作成した語彙リスト間で重複しなかった単語は上位 500 語では 120 語、6000 語では 504 語あった。

評価用ペアをみると、表 9 の評価例ように、高頻度語リストのみに出現した単語には「野党」、「海水浴」などジャンルや時期に偏りがあり、領域が狭いと評価されたものがあったのに対し、相乗平均で作成した語彙リストのみに出現した単語には偏りがなく領域が広いと評価されたものが多かった。

	単純頻度			相乗平均			判定不可能		
	頻度	領域	総合	頻度	領域	総合	頻度	領域	総合
500	10	4	3	34	46	39	6	0	8
6000	4	5	4	35	41	42	11	4	4

表 8：単純頻度−相乗平均の比較結果

上位 500 語

単純頻度	支持	野党	不安
相乗平均	立つ	上回る	女子
頻度/領域	相乗平均/相乗平均	比較不可/相乗平均	単純頻度/単純頻度
総合	相乗平均	相乗平均	単純頻度

上位 6000 語

単純頻度	スパイ	海水浴	組
相乗平均	遂行	断固	民放
頻度/領域	相乗平均/相乗平均	相乗平均/相乗平均	単純頻度/単純頻度
総合	相乗平均	相乗平均	単純頻度

表 9：単純頻度−相乗平均の評価例

5.3.考察

これまでの実験結果を踏まえ、「ニュースの基本語彙」の抽出に適した指標について考察した。

まず、表 4,表 6,表 8 にある各比較指標の総合評価を比較する。上位 500 語の評価では重み付き KL 情報量(式(3))の評価が最も良かったが、上位 6000 語の評価では低頻度語による悪影響が出てしまった。上位 500 語と 6000 語を合わせて評価した場合、相乗平均(式(2))の評価が最も良かった。

以上より、今回の実験からは、ジャンルと時期毎の単純頻度の相乗平均が、基本語彙として適切な単語を抽出できる指標だと考えられる。

6. まとめ

本稿では、ニュース中の頻度が大きく、領域が広い語彙を「ニュースの基本語彙」と定義し、それを抽出するのに最適な統計手法を求めることを目的とした。ただし、本稿における「頻度」、「領域」とは人間の内省によって決まるものとした。

そのため、「主観的頻度の大きさ」と「主観的領域の広さ」の 2 つの基準を用意し、人間の内省により近い統計指標を実験的に検討した。検討した指標は、文書頻度、重み付き KL 情報量、ジャンルと時期ごとの単語の頻度の相乗平均である。

その結果、頻度が大きく、時期やジャンルに偏らないで、領域が広いと人が感じる「ニュースの基本語彙」を抽出できる指標は相乗平均であることが分かった。

今後は、複数人の評価者による評価を実施して、より信頼性を高めていきたい。また、実際に「ニュースの基本語彙」を作成し、本稿で検討した指標の効果を確認したい。

文　献

[1] 田中 英輝, 美野.やさしい日本語によるニュースの書き換え実験.自然言語処理研究会, Vol.2010-NL-199 No.11, 2010

[2] 田中 彰夫.基本語彙と基本語.「日本語学」特集テーマ別ファイル語彙 1,pp.35-45,2005

[3] 国立国語研究所.日本語教育のための基本語彙調査.1984

[4] 凡人社.日本語能力試験出題基準改訂版.2006

[5] 内山,中條,山本,井佐原.英語教育のための分野特徴単語の選定尺度の比較.自然言語処理, Vol.11,No.3,pp.165-198,2004