擬似キーワード相関法による重要キーワードと重要文の抽出

亀田 雅之
(株) リコー・研究開発本部・情報通信研究所
kameda@ic.rdc.ricoh.co.jp

概 要

キーワード候補 (擬似キーワード) 間の部分文字列の重複に基づき、日本語文書中のキーワードやキーセンテンスを重要度付きで抽出する。

キーワードの重要度は、部分文字列の重複を利用した修正単語頻度及び単語長と字種から概算予測した構成単語数により評価する。また、文ごとに含まれるキーワード候補群同士の部分文字列の重複の総和により得た文間の関連度に基づく文マトリクスを利用して、文の重要度評価を行う。新聞記事や論文等の文書に有効で、重要度に基づいた適切なサイズの重要キーワードリストや抄録の作成や、関連度に基づいた関連文を提示する機能を実現できる。

1 はじめに

我々は、日本語文書の「読み」の支援として、軽量・高速な日本語解析系 QJP[1] を利用した日本語文書読解支援系 QJR[2] の機能を探っている[1]。QJR の4つの支援機能[2]のうち、Screening 支援では、キーワードが文書の対象や分野を端的に類推する手掛かりになることに着目し、簡易判定したキーワード候補群を提示して文書のふるい分け支援として利用し、Skip Reading 支援では、文書中の重要文を識別表示することで、飛ばし読みを支援することを狙った。

我々は、このキーワードと重要文を、文書検索システムでの検索式に対する検索文書の適合度や、検索文書をキーワード群や抄録等で内容を縮約してユーザに提示する支援機能にも有用であると注目している。特に、キーワードや文が重み付け (重要度付与) されているとこれらの扱いの可能性が拡がる。

キーワードは、従来、文書を検索するためのキーとして付与されてきた。自動的なキーワード抽出としては、構成単語や形態素のキーワード性を分析する方法 [3] があるが、一般には、文書内の単語の単語頻度が重要な指標とされている [4]。しかし、単語頻度では、同一の単語の出現がないと計数されない。一方、QJR では、複合語は、それだけ特殊性/専門性が高く、文書を特徴付けると考え、単語長に着目し、処理コストの小さい方法として QJP による形態素解析結果の品詞と単語長に基づいて文書中のキーワード候補 (擬似キーワードと呼ぶ；図1) を判定した[3]。

重要文抽出に関連する抄録の手法としては、接続詞等の手掛かり語により論理的な構成を分析し、重要な文を判定したり [5]、キーワードを多く含む文を抽出することが考えられる。前者は、文脈解析的側面が強く、難しい。この他[4]に、キーワードや抄録生成については、単語間の共起 [6] や分類語彙表に基づく関連性 [7] 等を利用した研究がある。

本稿では、日本語文書中のキーワードや重要文 (キーセンテンス) を重要度付きで抽出するために、擬似キーワード間の部分文字列の照合に基づいた擬似キーワード相関法と呼ぶ手法を主体にした方法を提案する。

通常兵器関連の工業製品
輸出規制が始動
4ヶ国対象

通常兵器の部品や加工機械に転用できる工業製品の輸出規制が二十日、日本でも始まった。英米などの主要先進七カ国（G7）の合意に基づいた規制であり、イラクなど四カ国を対象にして、対共産圏輸出規制委員会（ココム）のリストを準用する。G7は既に対象となる品目、国を広げるための話し合いを始めており、冷戦終結で変わる新たな輸出規制に発展しそうだ。

規制対象となる国は、イラン、イラク、リビア、朝鮮民主主義人民共和国（北朝鮮）の四カ国である。北朝鮮は、既に共産圏として特定地域に指定されているため、新たに追加されるのは三カ国である。また、イラクは経済制裁で終結管理が取られている。今回の措置で輸出に大きな変化が出るのはイランとリビアの二国になりそうだ。

輸出貿易管理令などに基づいて定められたコンピュータや工作機械などの規制品目を輸出する時には、通産省に許可申請を行う。その際、特定地域に指定されている国に対しては、明らかに民生用とわかる場合でなければ許可が下りず、事実上、規制品は輸出できない。

［文書出典：朝日新聞 1993年1月21日］

図1．ＱＪＲでの擬似キーワードの識別表示

[1] 我々のアプローチは、深い意味理解や内容理解には至らず、表層から形態素/構文レベルでの比較的軽い処理での実現を目指す

[2] 1.Screening(文書のふるい分け) 支援、2.SkipReading(飛ばし読み) 支援、3.Skimming(走り読み) 支援、4.Analytic Reading(分析的な読み) 支援

[3] 尚、単語頻度の考慮は必要であり、Screening 支援では、擬似キーワードを原文と同じ位置で抽出表示する他に、単語頻度ごとに単語長による評価で簡単に順序付けした一覧表で示す方法を示した

[4] [2] では、重要文の簡便な判定方法として Skip Reading と呼ばれる手法で取られている文書や段落の主意を表しやすい位置にある段落や文 (先頭/末尾段落、段落内の先頭/末尾の文) に着目する経験的な方法を紹介した

2 擬似キーワード相関法

2.1 擬似キーワード

文書におけるキーワードを文書内容の対象や分野を示す語[5]と考える。 QJRでは、数名詞や機能性の形式名詞や副詞名詞を除いた2文字以上の名詞をキーワード候補(擬似キーワード)とした。2文字以上としたのは、一般に1文字からなる和語名詞は、キーワードとはなりにくいという推測に基づく。本稿での擬似キーワードも同じ基準で選ぶものとする(図1)。

2.2 擬似キーワード相関法

キーワードの評価において単語頻度を指標とするのは、出現数の多い単語がキーワード性を高めているという観点である[6]。しかし、たとえば、図1の文書中のキーワード候補(擬似キーワード)の「輸出規制」、「対共産圏輸出統制委員会」、「規制」、「規制対象」、「輸出」、「輸出貿易管理令」といった出現単語については、「輸出」や「規制」という共通する構成単語がお互いに関連し、キーワード性を高め、文書を特徴付けていると見ることができるにもかかわらず、単語頻度だけではこうした寄与が見逃されてしまう。

擬似キーワード相関法[7]の基本的なアイデアは、こうした寄与を考慮することにある。擬似キーワード相関法では、上記の共通の構成単語の検出を単語間の部分文字列の照合で代替し[8]、その程度を単語間の関連度とみなし、また、以降で示すように単語頻度の修正として加味するものである。

擬似キーワード相関法の定式化として、擬似キーワード W_iの W_jに対する関連度 RW_{W_i}を次のように与える。

$$RW_{W_i}(W_j) = CW(W_i, W_j)/LW(W_i)$$

$CW(W_i, W_j)$: W_iと W_jの重複文字列長

$LW(W_i)$: W_iの文字列長

上記の例では、「輸出規制」と「対共産圏輸出統制委員会」との組では、CW(輸出規制, 対共産圏輸出統制委員会) が3(「輸出」と「制」の重複)であるので、$RW_{\text{輸出規制}}$(対共産圏輸出統制委員会) と $RW_{\text{対共産圏輸出統制委員会}}$(輸出規制) は、各々の文字列長で除して3/4と3/11となる。

さらに、この関連度を、文や段落、文書全体といった言語単位に含まれる擬似キーワード群に拡張する。即ち、擬似キーワード群 G_iの G_jに対する関連度 $RG_{G_i}(G_j)$ を次のように与える。

$$RG_{G_i}(G_j) = CG(G_i, G_j)/LG(G_i)$$

$$CG(G_i, G_j) = \sum_{W_n \in G_i, W_m \in G_j} CW(W_n, W_m) \quad [9]$$

$$LG(G_i) = \sum_{W_n \in G_i} LW(W_n) \quad [10]$$

3 キーワードの重み付け

キーワードの重み付けでは、従来の拡張として、前節の擬似キーワード相関法に基づく修正単語頻度に、予測構成単語数という指標を加え、擬似キーワードの重みを評価する。

3.1 修正単語頻度

修正単語頻度は、上記の擬似キーワード相関法に基づいて複合単語の構成単語レベルでの単語重複を単語頻度に加味して補正したもので、従来、キーワード抽出で利用されていた単語頻度に代わるものである。

修正単語頻度は、次のような類推に基づく。たとえば、4文字単語「輸出規制」は、「対共産圏輸出統制委員会」との3文字の重複によって、単語頻度を3/4だけ増やすと考える。一方、11文字単語「対共産圏輸出統制委員会」は、3回出現する「輸出規制」と3文字の重複があることから、$3/11 \cdot 3 = 9/11$ だけ単語頻度を増やすとする。同様に他の擬似キーワードのすべての寄与を合わせて単語頻度に反映する。

これを上記の擬似キーワード相関法の定式化の拡張として扱う。即ち、擬似キーワード相関法で対象とする擬似キーワード W_iの文書全体の擬似キーワード群Dに対する関連度 $RG_{W_i}(D)$ をもって、修正単語頻度 $FW(W_i)$ とする。即ち、

$$FW(W_i) = \sum_{W_j \in D} CW(W_i, W_j)/LW(W_i).$$

これによれば、「輸出規制」の単語頻度は3だが、文書中の他の擬似キーワードとの重複の寄与により、FW(輸出規制) は26/4=6.5となる。また「対共産圏輸出統制委員会」の単語頻度は1だが、文書中の他の擬似キーワードとの重複の寄与により、FW(対共産圏輸出統制委員会) は24/11=2.2となる[11]。

[5] キーワードは「何」について述べられているかを示す語であって、「どうした」という内容を示すものではない

[6] 対象文書群全体での頻度との比をとった方がよい指標となる

[7] 擬似キーワードに限定するものではないが、擬似キーワードに適用することで効果を上げていることからこの名称を与えた

[8] 簡便な部分文字列の重複の計数によるのは、形態素解析系とした用いたQJPが複合語の分割を行なわないということによる

[9] G_iと G_jに含まれる擬似キーワード同士の重複文字列長の総和

[10] G_iに含まれる擬似キーワードの文字列長の総和

[11] $FW_{W_i}(S)$ は、W_iと W_jが一致しない場合は、$CW(W_i, W_j)$ に0を与えるとすると、通常の単語頻度となる

3.2 予測構成単語数

QJR では、単語が複合しているということは、それだけ単語の意味の限定が強まるとして、単語長をキーワード性の判定に用いた。この考え方を一般化し、単語長に代え、構成単語数という指標を考える。

擬似キーワード W_iの構成単語数として、単語長とその単語の表記文字種に応じた1構成単語の平均長から概算予測した予測構成単語数 $NW(W_i)$ を用いる[12]。

$$NW(W_i) = LW(W_i)/lw(T(W_i))$$

$lw(T)$：文字種 T からなる構成単語の平均長

$T(W_i)$：W_iの表記文字種

たとえば、lw(漢字) = 2 なら、前記の例では、NW(輸出規制) は 4/2=2、NW(対共産圏輸出統制委員会) は 11/2=5.5 となる[13]。

3.3 擬似キーワードの重み付け

上記の修正単語頻度と予測構成単語数は、いずれも値が大きい程、キーワード性が高くなると考えられる。このことから、擬似キーワード W_iの重み付けの評価式 $WW(W_i)$ の1例として、c_f, c_nを重み係数として、次のように2要素の線形和を考える。

$$WW(W_i) = c_f \cdot FW(W_i) + c_n \cdot NW(W_i)$$

3.4 実験

上記評価式の2つの重み係数の何組かの組合せ試行のうち $c_f = c_n(= 10)$ の場合に、下記表1に示す再現率/正解率が比較的良い値を示した。図2に、その係数での図1文書の擬似キーワードに対する重み付けの例を示す。尚、ここでは、lwは統計値でなく、経験的な概数 [lw(漢字) = 2, lw(カタカナ) = 5] を用いた。

表1は、177 新聞記事ごとに人手で付与した上位10キーワードに対し、単語頻度のみ、修正単語頻度のみ、予測構成単語数のみ、及びこれらの線形和によって各々重み付けされた上位の擬似キーワードに基づく部分一致照合での再現率/正解率を示したものである。単語頻度に対しての修正単語頻度及び予測構成単語数の単独の重み付けの再現率/正解率が高く、さらに、これらの線形和による重み付けの再現率/正解率がさらに高いことが確認できる。

尚、他の単語に完全に含まれる単語は、上位キーワード抽出の際にリストから除いた (cf. 最長語併合 [6])。

[12] ここでは、形態素解析を行う QJP が複合名詞を分割しないことからこうした方法をとる

[13] 複数の表記文字種からなる複合語（例「擬似キーワード相関法」等）は、その文字種ごとの部分の予測構成単語数の和とする

4 文の重み付け

文の重み付けでは、各文を文中の擬似キーワード群で代表し、擬似キーワード相関法により文間の関連度を求め、関連度に基づき文の重要度を評価する。

4.1 文相関法

文内の擬似キーワード群に擬似キーワード相関法を適用する (特に、文相関法と呼ぶ)。文 S_iの S_jに対する関連度 $RS_{S_i}(S_j)$ を次のように与える。

$$RS_{S_i}(S_j) = CS(S_i, S_j)/LS(S_i)$$

$$CS(S_i, S_j) = \sum_{W_m \in S_i, W_n \in S_j} CW(W_m, W_n) \quad [14]$$

$$LS(S_i) = \sum_{W_m \in S_i} LW(W_m) \quad [15]$$

4.2 文の重要度の指標

$RS_{S_i}(S_j)$ 自体は、文間の関連度を示すものであるが、これを用いて、文の重要度の指標として、次のような指標を導入する。

平均関連度：$ARS_{S_i} = \sum_{S_j \in D, i \neq j} RS_{S_i}(S_j)/n'$

カバレージ：$CRS_{S_i} = \sum_{S_j \in D, i \neq j} \delta(RS_{S_i}(S_j))/n'$

n'：文書内文数−1　$\delta(x) = 0(x = 0), 1(x \neq 0)$

平均関連度は他の文との関連度の平均値であり、カバレージは他の文とどの程度広く関連しているかを示し、各観点での文の重要度に関わる。

[1]	85	(3)	輸出規制	[10]	40	(3)	イラク
[2]	77	(1)	対共産圏輸出統制委員会	[13]	35	(1)	加工機械
[3]	75	(1)	規制対象	[13]	35	(1)	禁輸措置
[4]	68	(1)	規制品目	[13]	35	(1)	許可申請
[5]	59	(1)	朝鮮民主主義共和国	[13]	35	(1)	工作機械
[5]	59	(1)	輸出貿易管理令	[17]	30	(2)	G 7
[7]	47	(1)	通常兵器関連	[17]	30	(1)	話し合い
[8]	45	(1)	主要先進七カ国	[17]	30	(1)	経済制裁
[9]	42	(2)	北朝鮮	[17]	30	(2)	リビア
[10]	40	(2)	特定地域	[17]	30	(2)	イラン
[10]	40	(2)	工業製品	[17]	30	(1)	冷戦終結

※ [10] 40 (2) 工業製品：[順位] 重み (単語頻度) 擬似キーワード

図2. キーワードの重み付け抽出 (上位20位)

表1. 修正単語頻度, 予測構成単語数による
キーワード抽出の評価 (再現率/正解率)

＼抽出数 (平均数)	上位5 (5.00)	上位10 (9.99)	上位20 (19.02)
単語頻度 Fw	27.9%/46.7%	48.8%/42.2%	73.5%/35.9%
修正単語頻度 FW	31.5%/55.7%	53.1%/48.6%	76.8%/38.6%
予測構成単語数 NW	33.7%/54.8%	54.7%/46.8%	78.5%/38.5%
Fw+NW	34.8%/57.1%	56.9%/48.6%	79.2%/38.7%
FW+NW	35.6%/60.3%	57.5%/52.0%	79.5%/39.6%

※対象文書：新聞記事177　※正解評価方法：部分一致照合
※正解キーワード：記事ごとに人手でつけたキーワード上位10個
※再現率：正解キーワード中、抽出キーワードで照合できた割合
※正解率：抽出キーワード中、正解キーワードで照合できた割合

[14] S_iと S_jに含まれる擬似キーワード同士の重複文字列長の総和
[15] S_iに含まれる擬似キーワードの文字列長の総和

4.3 実験

図3に、文間の関連度 $RS_{S_i}(S_j)$ を要素とする文マトリクス及び上記2指標値とそれらの積/和値を適宜、正規化して示した。第n行には第n文の他の文に対する関連度が示されている。見出し文は局所的に強い関連なのに対し、第5,7文は全体に広く関連している。

図4は、2指標値の積 $ARS_{S_i} \cdot CRS_{S_i}$ を重要度とした場合の上位文の抽出例である。経験的に重要と見られる見出しや先頭/末尾の段落や文の選択 [Skip Reading 手法] に近い。図5では、関連度により第1文の見出しの関連文を抽出表示した。短い見出し文の内容を敷衍する本文中の説明文が抽出されている。

5 まとめと今後の展開

擬似キーワード相関法は、文書の部分/全体の内容をその中に含まれる擬似キーワード群で代表させた上で、部分文字列照合手段により相互の関連度を得る手法である。本稿では、擬似キーワード相関法を提案し、これをベースに、擬似キーワードと文書全体の擬似キーワード群により修正単語頻度、文内の擬似キーワード群同士で文間の関連度を導き、キーワードと文の重み付けへの応用を示した。充分な評価には至っていないが、実験によりその効果を確認した[16]。また、文のカバレージや関連文といった新しい観点を示した。

本手法は、表層レベルの簡易なアプローチながら、文マトリクスに見られるような文脈レベルの手掛かりを与える。さらに、段落や章・節といった単位への拡張も可能である。

適切なキーワードや重要文の抽出には、従来の手法([3][5][6][7] 等) との組合せも検討するべきであるが、本手法は、単独でも小コストで高い効果が見込める。特に、論文や新聞記事等の漢語や外来語が多い文書に有効であると考えられる。

今後は、これらを利用して、QJR の支援機能を改善・拡張する一方、文書検索での検索式に対する検索文書の評価や文書の縮約提示 (図6[17]) 等に利用する予定である。また、重み付けや抄録の適切な評価法の検討や大規模な評価実験の実施も今後の課題となる。

謝辞 キーワード評価のためのキーワードデータ及びツールを提供してくれた小川泰嗣研究員に感謝する。

[16] キーワードの重み付けでは、予測構成単語数という単純な指標の有効性も確認した

[17] 擬似キーワード及び文の各重み付けを用いた縮約例である。重み値の閾値を変更することで縮約サイズを変更できる

```
[ i]: ..RSi(j)..  < ARSi , CRSi | Ai*Ci, Ai+Ci>   文の冒頭

[ 1]:*00800000000 < 7{11}, 9{12}| 0{12}, 8{12}> 通常兵器関連の
[ 2]:0*0777300373 <33{#1},63{ 5}|21{#1},48{ 3}> 輸出規制が始動
[ 3]:00*0AAA00000 <27{ 2},27{ 9}| 7{ 7},27{ 9}> ４ヶ国対象
[ 4]:420*22100131 <14{ 4},72{ 3}|10{ 3},43{ 5}> 通常兵器の部品
[ 5]:0111*2210111 < 9{10},90{#1}| 8{ 5},50{ 2}> 英米などの主要
[ 6]:02124*200131 <16{ 3},72{ 3}|11{ 2},44{ 4}> Ｇ７は既に対象
[ 7]:011132*21111 <12{ 6},90{ 1}|10{ 3},51{#1}> 規制対象となる
[ 8]:0000305*0004 <10{ 8},27{ 9}| 2{10},19{10}> 北朝鮮は、既に
[ 9]:00002020*200 < 4{12},27{ 9}| 1{11},15{11}> また、イラクは
[10]:020222302*20 <12{ 6},63{ 5}| 7{ 7},37{ 7}> 今回の措置で輸
[11]:0202221001*2 <10{ 8},63{ 5}| 6{ 9},37{ 7}> 輸出貿易管理令
[12]:02022223004* <14{ 4},63{ 5}| 8{ 5},38{ 6}> その際、特定地

 j :123456789ABC          ※ A,B,C : 10,11,12  ※ { }内 : 順位
```

図3. 文マトリクス

```
.....
輸出規制が始動
.....

  通常兵器の部品や加工機械に転用できる工業製品の輸出規制が二十日、日本でも始まった。.....
.....
.....Ｇ７は既に対象となる品目、国を広げるための話し合いを始めており、冷戦終結で変わる新たな輸出規制に発展しそうだ。
  規制対象となる国は、イラン、イラク、リビア、朝鮮民主主義共和国（北朝鮮）の四カ国である。.....
.....
                :
.....
```

図4. 重要文の抽出 (上位4文)

```
[1] 通常兵器関連の工業製品
→ [4] 通常兵器の部品や加工機械に転用できる工業製品の輸出規制が二十日、日本でも始まった。
```

図5. 第1文の関連文の抽出

```
【擬似キーワードーベスト8】
輸出規制／対共産圏輸出統制委員会／規制対象／規制品目／朝鮮民主主義共和国／輸出貿易管理令／通常兵器関連／主要先進七カ国

【重要文ーベスト2】
輸出規制が始動／Ｇ７は既に対象となる品目、国を広げるための話し合いを始めており、冷戦終結で変わる新たな輸出規制に発展しそうだ。
```

図6. 文書の縮約表示

参考文献

[1] 亀田雅之：怪盗・高速な日本語解析ツール『簡易日本語解析系 Q_JP』, 言語処理学会 第1回年次大会, 1995.

[2] 亀田雅之：日本語文書読解支援系ＱＪＲの検討, 情処学会 自然言語処理研究会報告 110-9, 1995.

[3] 小川泰嗣, 望主, 別所：複合語キーワードの自動抽出法, 情処学会 自然言語処理研究会報告 97-15, 1993.

[4] 久保田：キーワード抽出装置, 特開昭63-244259, 1987.

[5] 住田一男, 小野, 三池：対話的文書検索のための文書構造解析, 情処学会 自然言語処理研究会報告 97-11, 1993.

[6] 原正己：単語共起と語の部分一致を利用したキーワード抽出法の検討, 情処学会 自然言語処理研究会報告 106-1, 1995.

[7] 鈴木斎, 増山, 内藤：語彙的結束性に基づいた文章抄録法：情処学会 自然言語処理研究会報告 98-9, 1993.