放送ニュースを対象にした重要文抽出

加藤　直人[*1]　　浦谷　則好[*1*2]
*1 ＮＨＫ放送技術研究所
*2 現在，ＡＴＲ音声言語通信研究所
E-mail: katonao@strl.nhk.or.jp, uratani@slt.atr.co.jp

1　はじめに

インターネットの普及も手伝って文字情報があふれる中，自動要約技術が注目されており，自然言語処理システムの１つの機能として市販されているものもある．

このような自動要約のほとんどは，文章の中から重要な文を抽出するというものである．文の重要度は，キーワードの出現頻度，文章中での位置，手がかり表現などに基づいて計算している[奥村 99]．この中でも「キーワードの出現頻度」[Luhn 58]は，実現が容易であることから，重要文抽出に使われることが多い．「キーワードの出現頻度」による重要文抽出では，

（１）重要な単語（キーワード）は文章中で何度も出現する．

（２）重要文はキーワードを多く含む

と仮定し，要約率を満たす範囲で重要度の高い文を抽出する（あるいは，重要でない文を捨てる）という処理を行う．

確かに非常に長い文章ではこの仮定は有効であり，これを使った文の重要度は有効であることが期待できる．しかし，我々が対象にしている放送ニュースでは，１つの文章はそれほど長くはないので，前述の仮定が成り立たない場合も多い．したがって，「キーワードの出現頻度」に基づいた重要度では適切に文を評価できない場合もある．

本稿では放送ニュースの特徴を利用した重要文抽出手法について述べる．本手法では，ニュース文の１番目の文は文章全体の要約となっている文，すなわちリード文であるという特徴を利用して文の重要度を計算している．重要文抽出の際にはリード文をまず抽出し，それ例外の文の中から重要度が高い文を抽出する．

2　放送ニュースの特徴

放送ニュースは，文数が５～６文，単語数でも３００語程度であり，それほど長い文章ではない．さらに，キーワードの出現頻度をみても，頻度が高いものでも３～４回であり，突出して大きい値をとるわけではない．

また，各文に対してキーワードに基づいて文の重要度を計算してもそれほど大きな値の差となって表れない場合が多い．これは放送ニュースの特徴に起因する．放送ニュースでは始めにリード文があり，その次の文から具体的な内容の説明となっているという特徴がある．リード文は文章の要約となっているのでキーワードをほとんど含んでおり，「キーワードの出現頻度」による重要文抽出では必ず抽出できる．しかし，これに続く文では平均的にキーワードを含んでおり，「キーワードの出現頻度」による重要度ではその値に差が表れにくい．例えば，図１(a)の放送ニュースから重要文を抽出することを考えてみよう．「障害」，「解雇」，「倒産」という単語が複数回出現するので，「キーワードの出現頻度」による重要文抽出では，これらの単語が多く含まれる文１，文２，文５が重要度の高い文となる．要約率を５０％とした場合には，図１(b)のように文１や文２が抽出される．しかし，文２は表層的にも「去年１年間に企業の…身体に障害のある人は」という単語列が文１の前半と重複し，意味的にも同じ内容である．すなわち，図１(b)の要約は元の文章から落ちている情報量が大きい．要約としてはむしろ，リード文（文１）を抽出し，これと情報が重複していない文３や文４を抽出したほうが情報が落ちる割合が小さい．例えば，要約率を５０％とした場合には図２(c)のように文１，文３を抽出したほうがよい．

文１：去年１年間に企業のリストラや倒産などによって解雇された身体に障害のある人は、３０９３人と、前の年に比べて２倍近くに増加し、労働省では、規模の大きい企業で障害者の人達をもっと雇用してもらうことが出来ないかどうか検討を進めています。
文２：労働省によりますと、去年１年間にリストラや倒産などによって解雇された身体に障害のある人は、全国で３０９３人と３０００人を超え、およそ１６００人だった前の年に比べて２倍近くに増えました。
文３：また、障害のある人で仕事を探している求職者も去年の３月に初めて１０万人を超えてその後も増加を続け、現在は１１万人にのぼっているものと見られます。
文４：労働省によりますと、障害のある人を解雇した企業は規模の小さいところが多く、不況の影響で、リストラに踏み切り、その際、解雇するケースが目立つということです。
文５：このため、労働省では、経営的に体力のある規模の大きな企業に、障害者の人達をもっと雇用してもらうことが出来ないかどうか検討していて、具体策について、日経連と協議を進めています。

(a) 元の文章

文１：去年１年間に企業のリストラや倒産などによって解雇された身体に障害のある人は、３０９３人と、前の年に比べて２倍近くに増加し、労働省では、規模の大きい企業で障害者の人達をもっと雇用してもらうことが出来ないかどうか検討を進めています。
文２：労働省によりますと、去年１年間にリストラや倒産などによって解雇された身体に障害のある人は、全国で３０９３人と３０００人を超え、およそ１６００人だった前の年に比べて２倍近くに増えました。

(b) 重要文抽出結果（「キーワードの出現頻度」による）

文１：去年１年間に企業のリストラや倒産などによって解雇された身体に障害のある人は、３０９３人と、前の年に比べて２倍近くに増加し、労働省では、規模の大きい企業で障害者の人達をもっと雇用してもらうことが出来ないかどうか検討を進めています。
文３：また、障害のある人で仕事を探している求職者も去年の３月に初めて１０万人を超えてその後も増加を続け、現在は１１万人にのぼっているものと見られます。

(c) 重要文抽出結果（リード文の利用による）

図１　放送ニュースにおける重要文抽出の例

3　リード文を利用した重要文抽出.

本稿で述べる手法では，リード文を積極的に利用して重要文抽出を行う．すなわち，リード文は最も重要な文とし，リード文以外の文（以下では「本文」と呼ぶ）の中から，リード文と内容が重複するところが多い文を，要約率の範囲内で捨てるということを行う．このためには，ある文がリード文と内容が重複する部分を検出し，どのくらいの割合で重複しているかを求めておかなければならない．本手法では，以下で述べるように位置番号を使ってこれらを効率的に求めている．また，各文の重要度はリード文と単語対応がとれなかった割合に基づいて定義することにより，重複しない文の重要度を高くしている．以下ではこれらの詳細について述べる．

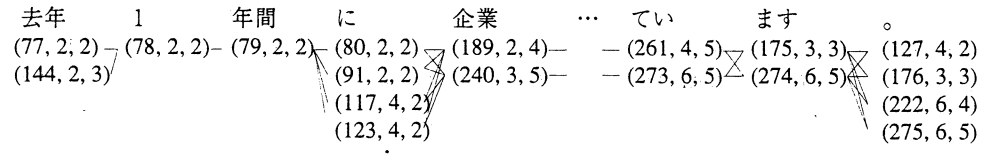
3．1　リード文と各文の単語対応

リード文と各文との単語対応をとる際に，単語対応はなるべく一つのまとまりとしてあったほうがよい．すなわち，連続した単語として対応がとれたほうがよい．あるいは，少なくとも同じ節内や文内で対応がとれたほうよい．本手法では以下のようにしてこれを実現している．

まず，文章中の各文の形態素解析を行い，各文を単語に分割する．文章の先頭から順に単語番号，節番号，文番号を付ける．これをまとめて位置番号と呼び，（単語番号，節番号，文番号）と表す．図１(a) の文章の例では図２(a) のようになる．

単語：	去年	1	年間	に	企業	の	リストラ	や	倒産	…
単語番号：	1	2	3	4	5	6	7	8	9	
節番号：	1	1	1	1	1	1	1	1	1	
文番号：	1	1	1	1	1	1	1	1	1	
(位置番号)：	(1,1,1)	(2,1,1)	(3,1,1)	(4,1,1)	(5,1,1)	(6,1,1)	(7,1,1)	(8,1,1)	(9,1,1)	

(a) 位置番号の例



(b) 単語対応の候補の例

(77, 2, 2) — (78, 2, 2) — (79, 2, 2) — (80, 2, 2) — (189, 2, 4) — — (273, 6, 5) — (274, 6, 5) — (275, 6, 5)

(c) 最適な単語対応の例

図2　リード文との単語対応の例

次にリード文中のすべての単語に対して，本文中の各文でも出現しているか否かを調べる．出現していれば，その単語の位置番号を記憶する．図1(a)のリード文では図2(b)のようになる．例えば，「去年」は文2にも出現しているので，その位置番号(77, 2, 2)を記憶している．ここで求められた位置番号が，リード文と本文中の各文との単語対応の候補となる．

最後に，リード文の文頭から文末までの単語対応の候補の中から，隣接する位置番号の距離の和が最短な経路を求める．隣接する位置番号の距離は，単語 w_i（i は単語番号）と単語 w_j（j は単語番号）との単語間の距離を表す関数（*distWordPost*）と，その文間の距離を表す関数（*distSentPost*）から式(1)のように定義した．

$$distPost(w_i, w_j) = \lambda_1 distWordPost(w_i, w_j) + \lambda_2 distSentPost(w_i, w_j) \quad (1)$$

（λ_1，λ_2（≧0）は定数，$\lambda_1 + \lambda_2 = 1$）

ただし，リード文中の単語が本文中に出現していない場合には，位置番号がないので，その1つ前の単語の位置番号を使うものとする．

単語間の距離を表す関数は式(2)で定義した．

$$distWordPost(w_i, w_j) = \begin{cases} log(j - i) & if\ j > i \\ log|j - i| \times penalty & otherwise \end{cases} \quad (2)$$

式(2)では，単語単号が文番号に比較して大きい値をとることもあるのでlogをとっている．

一方，文間の距離を表す関数は式(3)で定義した．

$$distSentPost(w_i, w_j) = \begin{cases} SentPost(w_j) - SentPost(w_i) \\ \quad if\ SentPost(w_j) \geqq SentPost(w_i) \\ (SentPost(w_i) - SentPost(w_j)) \times penalty \\ \quad otherwise \end{cases} \quad (3)$$

ただし，

$$SentPost(w) = SentNo(w) + \frac{PhraseNo(w) - 1}{PhraseNoMax(w)} \quad (4)$$

SentNo (w)：単語 w の文番号

PhraseNo (w)：単語 w の節番号

PhraseNo Max(w)：単語 w のが出現した文における節番号の最大値

式(3)では，同じ節中あるいは文中にある単語をを優先するためのものである．ここで，「節」とは今回は単純に「、」で囲まれた範囲としている．

式(2),(3)における*penalty*（≧1）は，番号が昇順である場合を優先するために，番号が逆転している場合にはペナルティを与えるためのものである．

最適な単語対応を求めるにはリードの文頭から文末までの位置番号間距離の和が最小となる組み合わせを求めればよい．これには動的計画法を使うことにより効率的に計算することができる．

3．2　文重要度

本文中の各文の重要度は，リード文と対応付けられた単語を多く含んでいないほど重要であるとし，式(5)のように定義した．

$$\begin{aligned} &ScoreSent(Sent) \qquad (5)\\ &= (1 - \mu_1 scoreWordCont(Sent) \\ &\quad - \mu_2 scoreWordFunc(Sent)) \times 100 \end{aligned}$$

$$scoreWordCont(Sent) = \frac{\text{文}Sent\text{の内容語の中でリード文と一致した個数}}{\text{文}Sent\text{の内容語の個数}}$$

$$scoreWordFunc(Sent) = \frac{\text{文}Sent\text{の機能語の中でリード文と一致した個数}}{\text{文}Sent\text{の機能語の個数}}$$

（μ_1，μ_2（≧0）は定数，$\mu_1+\mu_2=1$）

例えば図1 (a)の例で，パラメータの値を

$\lambda_1 = 0.1$（$\lambda_2 = 0.9$）

$penalty = 1.5$

$\mu_1 = 0.1$（$\mu_2 = 0.9$）

としたとき，本文中の各文の重要度は次のようになった．

文2：34.7

文3：95.7

文4：82.5

文5：59.0

重要度を比較すると，文3，文4が重要であることがわかる．また，実際の文章（図1 (a)）を見ると，これらの値はそれぞれの文の重要性に対する直観と一致していることがわかる．

5　おわりに

リード文を利用することにより放送ニュースから重要文を抽出する手法について述べた．今後は，本手法の評価をする予定である．また，「キーワードの出現頻度」に基づく手法との比較も行いたい．

今回導入した位置番号の中で，単語番号はそもそもは音声認識の言語モデルにおいて(n≧4)を利用するときに，連続している個所を発見するのに使っている[加藤99]．しかし，単語番号はその他にもいろいろ応用できることがわかっている．今後はこのような応用についても考えていきたい．

【参考文献】

[加藤99]加藤ほか「ニュース音声認識のための(n≧4)-gramを併用する言語モデル」電子情報通信学会研究報告，SP99-125, pp.55-60, 1999.

[Luhn 58] Luhn, H.P. "The automatic creation of literature abstracts, In IBM Journal for Research and Development.", 2(2), pp59-165, 1958.

[奥村99]奥村．難波「テキスト自動要約に関する研究動向」自然言語処理，6(6)，pp.1-26, 1999.