ニュース文を対象にした局所的自動要約手法

6R－9

加藤　直人
NHK放送技術研究所

1　はじめに

TVニュースの文章を自動的に局所的要約する手法について述べる．局所的要約とは原文の単語列を置換することによる要約をいう．局所的要約は文章の自動要約におけるプロセスの1つであり，字幕作成への応用なども考えられる．以下では，単純に要約と書いた場合には局所的要約を指すものとする．

自動要約には，どのような単語列をどのようなときに置換すればよいかという要約知識が必要となる．従来の要約知識は人手で収集していたためその数は少なく，実際に適用する際にも順次適用すればよかった．しかし，最近，要約知識を自動的に獲得する手法[加藤 98]が提案され，要約知識を大量に獲得することが可能となってきている．要約知識が増えると，複数の要約知識を適用できる場合があり，要約知識間の競合の問題が生じる．

本稿では，コスト付きの要約知識を使い，文全体として最適な局所的要約をするアルゴリズムについて述べる．

2　自動要約

2．1　要約知識

自動要約に必要な要約知識について簡単に述べる．要約知識は自動獲得しているが，詳細については文献[加藤 98]を参照されたい．

要約知識は，置換知識と置換条件の2つから構成されている．置換知識は原文の単語列をどのような単語列に置換するかを規定する知識である．例えば，次の例は連体助詞の「の」を省略するという置換知識である．

【置換知識の例】

「「の／体助」→ ϕ（省略）」

一方，置換条件とは置換知識の適用の良否を数値化したもの，すなわち置換知識のコストである．置換知識はその前後の単語列によっては適用の良否が決まる．例えば先の置換条件の例を使って，「日本の銀行」の「の／体助」を省略することはできない．そこで置換知識のコストは，置換知識の前後の単語列と，あらかじめ獲得しておいた置換条件との距離を用いて計算している．すなわち，i番目からj番目までの単語列 $w_{i,j}$ を，単語列 $x_{i,j}$ に置換するという置換知識のコストを $distsub(w_{i,j}\rightarrow x_{i,j})$ と表すと（1）式で定義される．

【置換知識のコスト】

$$distsub(w_{i,j}\rightarrow x_{i,j}) \quad (1)$$

$$= \begin{cases} g'(w_{i,j}\rightarrow x_{i,j}) & if\ \text{正例があるとき} \\ 0.0 & otherwise \end{cases}$$

ただし，

$$g'(w_{i,j}\rightarrow x_{i,j}) = (1.0-g_{low})\times g(w_{i,j}\rightarrow x_{i,j}, \text{正例}) + g_{low} \quad (1a)$$

（$g(w_{i,j}\rightarrow x_{i,j}, \text{正例})$ の定義は[加藤 98]）

$$g_{low} = 0.01 \quad (1b)$$

（1）式は，正例がある場合には g_{low}（=0.01）～1.0の値（$0.0 \leq g(w_{i,j}\rightarrow x_{i,j}1.0, \text{正例}) \leq 1.0$）を取り，0.0に近いほど置換することが可能であると定義されている．また，正例がない（適用される置換知識がない）場合には0.0を取る．

2．2　自動要約アルゴリズム

説明を簡単にするために，以下では1文を要約する場合を考える．複数の文にわたる場合には単純に連結すればよい．

今，原文をある要約率（要約文の文字数／原文の文字数）以下に要約したいとする．このとき，m（＝原文の文字数×要約率）文字以上の文字数を削除しなければならない．さらに，最適な要約であってほしい．ここで「最適な要約」とは，適用した置換知識のコストの和（置換コスト）が最小となる場合であると定義する．したがって，自動要約とは，m 文字以上の文字数を削減し，文頭から文末までの置換コストが最小のパス（最適パス）を求めることである．定式化すると，（2）式のようになる．

【局所的自動要約の定式化】

$$\underset{x\in X}{argmin} \sum_{x} distsub(w_{i,j}\rightarrow x_{i,j}) \quad (2)$$

$$X=\{(x_0, \cdots, x_{i,j}, \cdots, x_{,n}) \mid \sum (|w_{i,j}|-|x_{i,j}|)\geq m)\}$$

A new method of finding an optimal summarization of a TV news sentence using automatically-scored summarization knowledge.

Naoto Katoh
NHK Science and Technical Research Laboratories
1-10-11 Kinuta Setagaya-ku Tokyo 157-8510, Japan

（2）式の解を求めるアルゴリムについて説明する．アルゴリズムは図1のように3つのステップから成っている．このうち第1，2ステップが最適パスを求める部分であり，Soongらが連続音声認識のために提案している探索アルゴリズム[Soong 91]を応用している．このアルゴリズムでは，A* アルゴリズムに必要なヒューリスティック関数（現在ノードからゴールまでの評価関数の予測値）をあらかじめダイナミックプログラミング（DP）により求めておく．本手法では文字削減数と置換コストという2つの評価関数を用いているが，前者を計算する際にヒューリスティック関数を用いている．

第0ステップでは原文の形態素解析を行い，その結果に対して要約知識を適用する．置換知識による単語列も追加すると，図1のような単語のラティス構造とともに，置換知識のコストが計算される．

第1ステップでは，DPによりヒューリスティック関数の値を求めておき，解となる見込みのないパスを第2ステップにおいて枝刈りする際に使用する．文末から文頭（後向き）に探索し，文末から現在ノードkまでの最大可能な文字削減数（後向き文字削減数）$m_b(k)$を求める．現在ノードが開始ノードである場合（$m_b(0)$）が，この文に可能な最大削減文字数である．この値が所望の値mよりも小さいとき（$m_b(k)<m$）は希望の要約をできない旨出力して終了する．

第2ステップではA* アルゴリズムにより最適な解を探索する．文頭から文末（前向き）に探索し，文頭ノードから現在ノードkまでの文字削減数（前向き文字削減数）$m_f(k)$，および置換コスト$cost(k)$の組（$m_f(k)$，$cost(k)$）を計算する．文字削減数を計算する際に，前向き文字削減数と後向き文字削減数の和が所望の値mよりも小さい場合（$m_f(k)+m_b(k)<m$）には，このパスは最終的な解となりえないので枝刈りする．文末にいくにしたがい可能なパスの候補が増加していくが，このような枝刈りにより候補数を抑えることができる．

以上の説明では適用される置換知識がそれぞれ独立であるとした．しかし，「総理大臣→首相」のように，1度適用したら次回にも必ず適用しなければならない置換知識もある．これに対応するためにはパスごとに必須適用置換知識リストをもっておき，置換知識を適用するときにそのリストを参照する処理を本アルゴリズムに追加すればよい．

4　おわりに

与えられた要約率以下で，最適に要約する局所的要約手法について述べた．本アルゴリズムの第2ステップ中には，正例がない単語（例えば図1の「で」）に達したときに，条件（前向き文字削減数，必須適用置換知識リスト等）の同じ候補は，置換コストが最小でないパス（例えば，図1のパス2）も枝刈りするという改善を加えることも可能である．

今後は本手法を計算機上にインプリメントし，文章全体を局所要約する評価実験を行う予定である．

参考文献

[加藤 98]加藤直人：ニュース文要約のための局所的要約知識獲得とその評価，電子情報通信学会，言語理解とコミュニケーション研究会，NLC98-16, pp.7-14 (1998).

[Soong 91]Soong, F.K. and Huang E. :A Tree-Trellis Based Fast Search for Finding the N Best Sentence Hypotheses in Continuous Speech Recognition, ICASSP-91, pp.705-708 (1991).

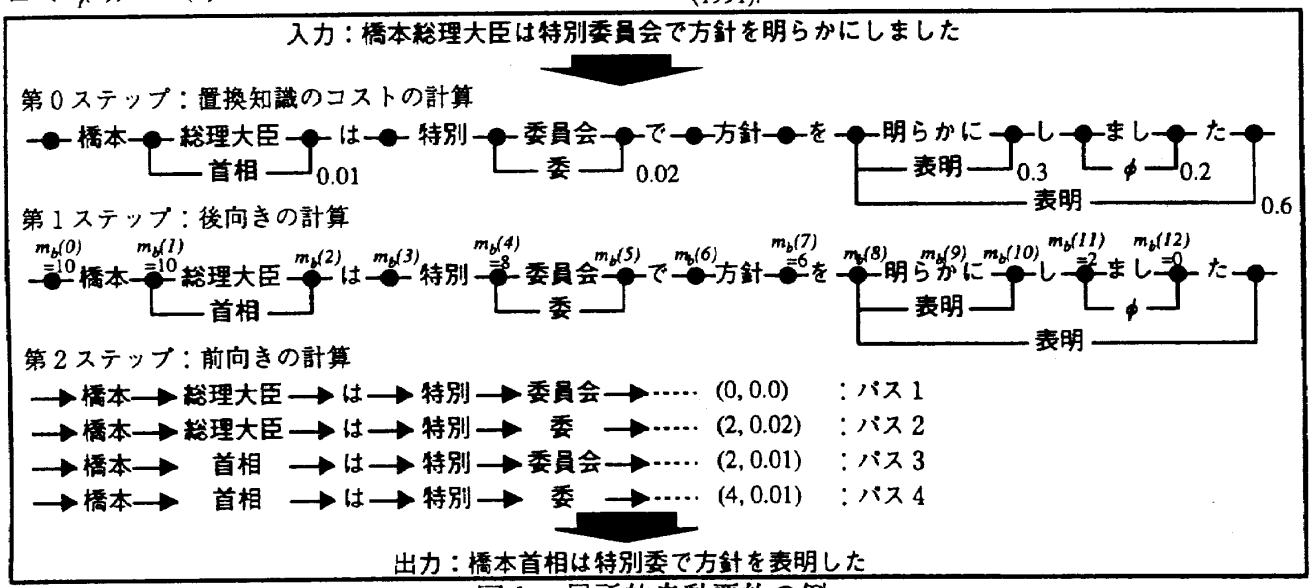

図1　局所的自動要約の例