3C−9

重要文抽出と不要語句削除による抄録作成方式

森本康嗣
（（株）日立製作所　システム開発研究所）

1．はじめに

Luhn[1]に始まる自動抄録(要約)システムに関する研究は、現在、談話理解に対する興味だけではなく、必要な情報を短時間で得ることを可能にするという実用的な価値のため、大学から企業まで広く行われている。本稿では、大量の文書からの情報のスキャニングを目的として、表層・統計情報を利用する重要文抽出法と英日機械翻訳システムの構文解析技術を組合せた抄録作成方式を提案する。

2．抄録作成方式の概要

本稿で提案する抄録作成方式は、

(1) 抄録候補文抽出ステップ

(2) 不要語句削除ステップ

という２つのステップからなることを特徴とする。

2.1 抄録候補文抽出ステップ

重要な文を抄録文の候補として抽出する。重要文抽出の手がかりとしては、次の二つを用いる。

(1) 重要な箇所であることを示す表層的表現

これは、"in summary","to be remarkable"などの、重要な箇所であることを示す表現（以下、重要表現と呼ぶ）であり、英文要約システムDIET[2]などで利用されている。

この手がかりの長所としては、

① 重要表現が含まれている文は、実際に重要である可能性が高く、重要ではない文を誤って抄録に採用する可能性が低い。

また、短所としては、

① 重要表現が使われていない重要文を検出することはできない。

② 同種の表現に、"for example","in other words"のように、例示、言い換えなど抄録には不要な文であることを示す表現(以下、不要表現と呼ぶ)も存在する。しかし、重要表現を含む文を抄録に採用する方法では、不要表現を利用できない。

③ システムのユーザの立場から言えば、ユーザが抄録の量を指定できる(例えば、「元の四分の一」、「A4一枚」など)ことが望ましい。しかし、重要表現を含む文を抄録に採用する方法では、抄録の量を制御することができない。

(2) キーワードの密度

Luhn[1]以来、多くの改良方式が提案されている。基本的には、高頻度語をキーワードとみなし、キーワードが高密度で含まれている文を重要と考える。

この方法の長所としては、

① 文を抄録に採用する際に、キーワード密度の閾値を変化させることにより、抄録の量を自由に制御できる。

② 文章の種類を問わず適用できる。

また、短所としては、

① 統計情報のみに基づいて文の重要さを評価するため、非重要文を抄録に採用したり、重要文が抄録から漏れたりすることがある。

そこで、この二つの手がかりを次のように組み合わせることにした。

(ステップ１) キーワード密度に無関係に、重要表現を含む文は抄録に採用し、不要表現を含む文は抄録に採用しない。これにより、キーワード密度が低い重要文が抄録から漏れたり、キーワード密度が高い非重要文が抄録に採用される可能性を減少させることができる。重要表現を含む文が予め定められた抄録候補文数を越えた場合は、ステップ１で処理を終了する。

(ステップ２) 重要表現・不要表現のどちらも含まない文については、キーワード密度法によって抄録候補文を抽出する。これにより、重要表現を含まない文を抄録に採用すること、ユーザが希望する量の抄録を作ることができる。

さらに、キーワード密度法において、文章が段落から構成されている点に注目した。段落は、意味的なまとまりの単位であり、文章の構造をある程度反映している。そして、段落単位で考えても、結論などの重要な段落や、具体例などの不要な段落が存在する。そこで、段落にキーワード密度に基づいて得点を付け、段落の得点に比例するように各段落から抽出する抄録候

An Abstracting Method based on Extracting Important Sentences and Deleting Unnecessary Words
Yasutsugu MORIMOTO
Systems Developement Laboratory, Hitachi, Ltd.

補文数を決定する。

2.2 不要語句削除ステップ

抄録候補文を出現順に並べて抄録を作成する際に、不要な語句を削除する。不要である条件として以下の二つを設定し、いずれかの条件を満たす要素を不要な要素と考える。

(1) 別の要素の同格要素であること

同格の要素は、冗長な情報だと考え削除する。

(2) キーワードを含まないこと

キーワードを含まない要素は、抄録候補文抽出ステップと同様に重要ではないと考え、削除する。

ただし、実験システムにおいては、不要な語句として削除するのは任意格要素に限定し、不要な要素でも必須格要素は残すことにした。これは、ベースとした英日機械翻訳システムが構文トランスファ方式を用いており、必須格要素を削除すると構文的に正しい文を生成することが困難なためである。

不要語句削除処理の効果として、

(1) 1文の長さを短くできるので、重要文全体を抽出して抄録を作るときよりも、抄録の量を小さくすることができる。逆に言えば、抄録全体の大きさが同じならば抽出される文の数を多くすることができるので、重要文が抄録から漏れる可能性を減少させることができる。

(2) 1文の長さが短くなるので、読み易さが向上する。

また、機械翻訳システムという観点からは、

(1) 翻訳すべき文数が減るので、トータルの翻訳時間が減少する。

(2) 翻訳処理の途中で解析木が簡単になるため、変換・生成処理での失敗が少なくなり、見かけ上翻訳精度が向上する。

3．評価

3.1 抄録候補文抽出ステップの評価

"Business Week"から3つの記事("March 20, 1989, pp.87-88.","June 13, 1988, pp.56-57","August 8, 1988, p43")を選び、提案方式および従来のキーワード密度方式によって、元の文書の1/2および1/3の量の抄録候補文を抽出する実験を行った。その際、3人の被験者に記事を読んでもらい、2人以上が重要だとみなした文を、重要文として評価基準に用いた。各方式における重要文抽出率およびノイズ混入率の平均値を以下に示す。ただし、

重要文抽出率=抽出された重要文数/全重要文数

ノイズ混入率

=(抄録文数-抽出された重要文数)/抄録文数

	従来方式		提案方式	
抄録の量	1/2	1/3	1/2	1/3
重要文抽出率[%]	62.9	43.8	71.1	51.1
ノイズ混入率[%]	72.6	71.0	68.9	65.1

提案方式によって、従来のキーワード密度方式より重要文抽出率が10%程度向上した。ただし、ここで重要文抽出率の向上に寄与したのは段落の重み付けの導入のみであった。これは、対象文書においては重要表現・不要表現が存在しなかったためである。

3.2 不要語句削除ステップの評価

"Business Week, March 20, 1989, pp.87-88."から1/2の文(30文)を取り出したのち、不要語句削除処理を行った結果を示す。

全文の語数	抄録候補の語数	削除された語数
1059	564	95

この文書に対しては、抽出された抄録候補文中の約2割弱 (95/564≒0.17)の語句を削除することができた。今後、大量の文書に対し、評価を行う。また、削除した語句の妥当性の評価も今後の課題である。

4．おわりに

本方式の特徴は、(1)重要表現・不要表現とキーワード密度法の組合せにより重要文抽出を行う点、(2)段落の重みの導入によりキーワード密度法を改善した点、(3)抽出された重要文から不要語句を削除し抄録量を減少させる点である。今後の課題としては、(1) 抄録精度の向上、(2) 抄録の一貫性・結束性の改善[3]が挙げられる。

参考文献

[1] Luhn, H.P:The Automatic Creation of Literature Abstracts, IBM Journal, vol.2, (1958).

[2] 石橋他:英文要約システム「DIET」, 情報処理学会第38回全国大会, (1989).

[3] 田中:抄録のための言語処理, 朝倉日本語新講座6 運用2・人文系研究のための言語データ処理入門, pp.1-41, 朝倉書店, 1983.