SVM による重要語選択に基づく要約システム ABISYS2004

田中 信彰(15801044)　広瀬 裕二(15801063)

原田研究室　指導教員　原田実　教授　　韓 東力　助手

1. はじめに

原田研究室では 2002 年度より、日本語文章から SAGE[1]によって意味解析をし、そこから得た格フレーム群を使い要約文を生成する ABISYS[2]を開発している。本研究では ABISYS を SAGE の新版に合わせて C#で全面的に改訂すると共に、いくつかの精度向上を図った。従来の ABISYS では重要語の選定においてマハラノビスの汎距離を用いて判断をしていた。本研究では重要語の選定にパターン識別法である SVM を用いた。また要約文生成において要約要素語の選定方法や要約要素語の用言の必須格の抽出の判断基準である閾値の算出方法を改善し、出力すべき要約要素語と重要語の関係を見直すことでより自然な要約文の生成を可能にした。

2. 重要語得点要素

本研究では、体言が要約文章の中心であると考え、体言（形態素）を重要語候補とする。得点は重要度合いの反復得点、文間得点、位置得点、意見得点、主題・焦点得点の 5 種類とし、それらの得点を SVM によって総合化することで体言に対する重要語得点を求める。以下にそれぞれの得点の求め方を示す。

2.1. 反復得点

ここで言う反復語とは、文章中の異なる文に 2 度以上出現する同一語・同意語・同カテゴリ語のことである。同一語とは見出しが一致する語、同意語とは見出しは異なるが概念 ID(語意)が等しい語、同カテゴリ語とは語と語の概念距離がもっとも近い語を指す。

2.2. 文間得点

文間深層格とは、隣接する 2 文間の意味的な関係を定義したもので 21 種類の格がある。この文間深層格と要約の関係を調査したところ、話の転換を示す inter-conversion 等一定の格を持つ文が要約に残りやすく、詳細化を示す inter-detail 等の一定の格を持つ文が残りにくいことが分かった。この調査を基に文を構成する体言に対して加点・減点する。

2.3. 位置得点

新聞記事では冒頭文が重要であるとされ、最終文にはテーマに関する今後の展開などが書かれる場合が多い。論説文においても冒頭文および最終文は筆者のまとめが書かれることが多いとされている。ABISYS では、第 1 文目と最終文を重要視し、それらの文中の体言に対して加点する。

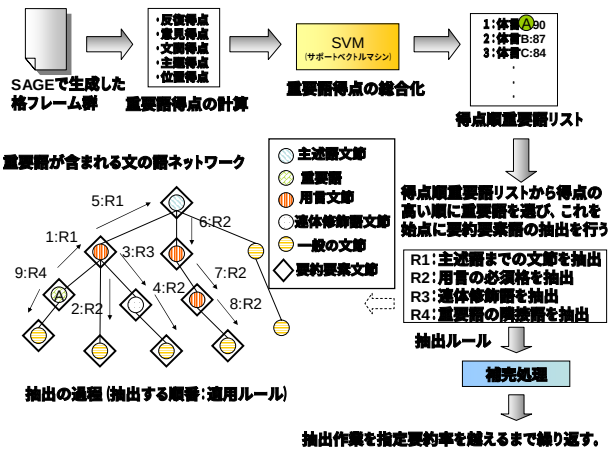

図 1 ABISYS による要約手順

2.4. 意見得点

論説文においては、筆者の主張・意見、希望がもっとも重視されるものである。そこで意見語を特定するための意見語句(「思う」など)を調査し、その語句と深層関係がある語、さらにその語とも深層関係がある語句に対して意見得点として加点する。これは意見語句から 2 以内の深層関係において筆者の意見が集中しているためである。

2.5. 主題・焦点得点

要約において、原文の主題や結論は重要であり、主題・焦点を特定できれば、原文の趣旨を読み取った要約を作成することが出来る。そこで ABISYS では、次のようなルールで簡易的に主題・焦点を決定し、それぞれの体言に対して加点する。(1)名詞 A＋助詞「は」「も」「が」「を」「に」「で」の形を探す。(2)この名詞 A の中で他の語句に対して「agent」「object」「a-object」「goal」という深層格で接続する名詞 B を特定する。(3)この名詞 B を主題･焦点語とし加点する。

3. 重要語句得点の総合化

従来の ABISYS ではマハラノビスの汎距離を用いて確率的な重要語句決定方法を採用していた、今年度はより精度を高めるため、パターン判別方法である SVM（サポートベクターマシン）を採用した。ABISYS では 5 変数で定義された空間、つまり「反復得点」、「文間得点」、「位置得点」、「意見得点」、「主題・焦点得点」の５得点からなる５次元空間において、 重要・非重要語

を分離する超平面を学習する。学習データとして用意した事例文章中の体言を人手で重要・非重要語の 2 つに判別しておく。これらのデータに対して SVM を適用し、判別式 g(x)=wx+b>0 の w と b を学習する。要約率に応じて、要約対象の文章中の全体言に対してこの総合得点 g(x)の高い順に重要語句とする。30 文章中 1887 個の重要語句候補を用いて SVM とマハラノビスとの比較実験を行ったところ表 1 のような結果となり、結果的に SVM の方がマハラノビスよりも正解率が高いことがわかった。

	正解数	重要語句数	正解率
SVM	1758	2669	65.9
マハラノビス	1472	2669	55.2

表 1 SVM とマハラノビスの実験結果

4. 自然で正しい日本語としての出力

抽出された重要語句から自然で正しい日本語としての要約文を出力するための方法を述べる。

4.1. 重要語句から要約要素語の抽出

重要語句からの要約要素語の抽出は 2 に示すように次の 4 つのステップで行う。(R1)重要語句から主述語までのパス上の語を要約要素語として抽出し、さらにこれが用言であればその必須格を、また連体修飾語を持てばその語も抽出する。(R2)重要語句と深層格で隣接した語句と「reason」「cause」「sequence」格で隣接する用言を抽出する。 (R3)抽出した要約要素語中に、「こと」「の」「という」「とおり」が含まれていた場合、これらの語は何か他の語を伴って意味が生ずると考えられるので、さらにその modifier 格の語を抽出する。(R4)なお、このような必須格や連体修飾語の抽出は新しく抽出された要約要素語に対しても条件に合うものがある限り抽出を繰り返す。

4.2. 必須格の抽出

本研究では、EDR 単語辞書中の全用言の各深層格 x の共起辞書内における出現割合を 0.01%刻みで頻度展開し、積分して上位 90%、80%、70%における出現割合の閾値 m90、m80、m70 を求め、用言 v の格 x の出現割合が閾値以上であれば、x を v の必須格とした。必須格である要素が用言の場合はさらにその必須格も抽出する。また用言の必須格があまりに少ない場合は、EDR 辞書におけるデータスパースと考え、不自然な文章となることを防ぐために、深層格でつながる要素をすべて抽出する。実験の結果、閾値として m90 を用いるのが、要約文の日本語としての正しさと要約箇所の適切さから、最も好ましい要約文を得られることがわかった。

5. 文を縮める技術

引用動詞の削除などを行い、意味的に重複するものを削除することにより、要約文を短くする。

5.1. 引用動詞の削除

引用動詞の削除は次の 4 ステップで行う。(1) 上位概念として「考える」 “30f878”や「思考する」“444dda”を持つ語 A を探索する。(2) 語 A が持つ object 格の語 B を探索する。(3) 語 B の品詞が用言であり、かつ語 B の助詞が「と」「ように」「とか」ならば、語 B と深層格でつながる語 A を引用動詞と判断し、語 A を削除する。(4) 語 A の引用節以外の深層格による宛先の語を削除する。

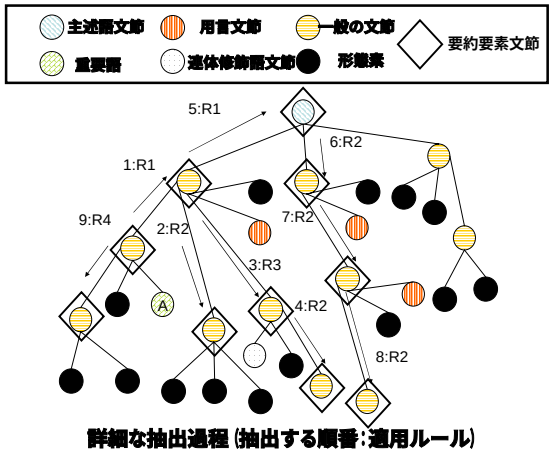

図 2 要約要素語の抽出

6. おわりに

本システムと従来のシステムによる要約と、人手による要約を主観評価と文の修正の度合いに関して評価したところ、従来の手法に比べ、要約品質、速度ともに向上した。よって本システムの有効性を示すことができたと言える。

参考文献

[1] 前澤敏之,面来道彦,上野雅和,韓東力,原田実:”意味解析システム SAGE の高精度化と概念グラフへの変換”,情報処理学会第 66 回全国大会論文集,6U-05,第 2 分冊 pp.177-178(2004.3).

[2] 野口貴,韓東力,原田実:”反復語句・必須格・文間深層格を考慮した要約システム ABISYS”,情報処理学会第 66 回全国大会論文集,6U-04,第 2 分冊 pp.175-176(2004.3).