2D-2

章節構造を持つ文書にも対応した自動要約システム ABISYS2010

原田　実†
青山学院大学理工学部情報テクノロジー学科†

西尾　公秀‡
青山学院大学大学院理工学研究科知能情報コース‡

1. はじめに

近年，膨大なテキスト情報から必要な情報のみ抽出する必要性が増大し，自動要約の重要性が増している。

原田研究室では，日本語文の意味解析システム SAGE[1]から出力される意味グラフを用い要約文を生成するシステム ABISYS[2]を開発してきた。しかし，現状では長くても数百字程度の新聞記事や論説文には有効であるが，論文や特許文などの章節構造をもった長文からなる文書の解析では文章の前半と後半に要約結果が集中してしまうなど，良い要約結果を得られない。そこで本研究では，センタリング理論に基づいて文書を談話構造解析するシステム DIA[3]を用いて，文全体の話の焦点の推移を考慮した重要語得点の導入を行い，章節構造に基づいた論文などの長文の要約手法の開発を行う。

2. 自動要約システム ABISYS の処理の流れ

ABISYS の処理の流れを図 1 に示す。入力した文章を SAGE を用いて意味解析し，さらに DIA を用いて談話構造解析する。その結果から各体言を重要語候補とし，表層情報，深層情報に基づいて重要語得点を計算する。それを元に，ランキング SVM[4]を用いて識別関数を計算し，得点が大きいものから順に，要約率を満たすまで要約文を生成する。

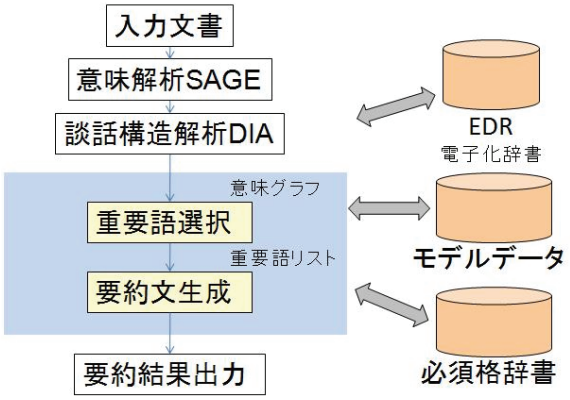


図 1 ABISYS の処理の流れ

Automatic summarizing system ABISYS2010 for documents with chapter structure
†Faculty of Science and Engineering, Department of Integrated Information Technology, Aoyama Gakuin University
‡Graduate School of Science and Engineering, Intelligence and Information Course, Aoyama Gakuin University

3. 重要語得点

ABISYS では従来より文中の体言ごとに重要語候補とし，反復得点，主題得点，位置得点，意見得点，文脈得点，属性得点，固有表現得点の 7 得点を計算している。

3.1 従来の重要語得点

以下に従来の重要語得点を示す。

・反復得点
異なる文に二度以上出現する重要語候補と同一語句，同義語句，同カテゴリ語句を抽出し，意味的な反復回数に応じて反復得点として加点する。

・主題得点
主題や焦点を特定するため，重要語候補の助詞や深層格によって主題得点として加点を行う。

・位置得点
文の出現位置によって位置得点として加点する。解析対象が論文の場合は文全体における文の位置ではなく，章，節，項内での文の位置が重要度に影響する。そこで，各章ごとに先頭 3 文と後方 2 文に対して得点を与える。

・意見得点
意見語句(「思う」など)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。

・文脈得点
隣接する 2 文間の意味的な関係を定義した文間深層格 21 種類について，各文間深層格に応じて加点・減点を行う。

・属性(モダリティ)得点
モダリティとは，筆者の命題に対する主観的認識や発話態度を表すものである。文のモダリティによって，属性得点に加点を行う。

・固有表現得点
SAGE によって重要語候補が固有表現と判定された場合，固有表現得点に加点を行う。

3.2 新しく導入した得点

本研究ではこれに談話構造得点，見出し得点を加えた 9 得点を重要語得点とする。

3.2.1 談話構造解析

談話構造解析システム DIA は，図 2 に示すような焦点の推移を表す談話構造木を自動構築し，文間の接続関係を決定する。

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

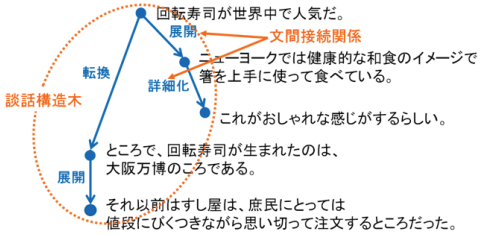


図 2 DIA による談話構造木

3.2.2 談話構造得点

話の焦点の流れを推定するために，談話構造解析結果に基づいた得点を導入する。DIA によって得られる談話構造木において，先頭の文から最後の文までのパス上にある文には焦点の移り変わりが記録されていると考え，ここに現れる文節に対して談話構造得点を与える。論文では，1 つの章を単位とした談話構造解析結果を用いる。

3.3 章節構造を用いた解析

3.3.1 章節構造の認識

論文の要約において，全体の流れを踏まえたまとまりのある要約を生成するためには章節構造が重要な手掛かりになる。そこで，論文の解析では章節構造に基づいた重要語得点を用いる。章節構造の解析は見出しに基づいて行う。見出しは正規表現によって判定し，章，節，項の順に認識する。

3.3.2 見出し得点

論文において，見出しはその章の中で重要な役割を持つ。そこで，重要語候補が属する章・節・項それぞれの見出しに含まれる名詞と重要語候補が一致した場合，見出し得点として加点する。

3.3.2 要約率による処理

要約率が 30%以上であれば章の見出しを，50%以上であれば，すべての見出しを要約結果に残す。要約率が 30%以上であれば，談話構造解析は章を単位として行うが，それ以下の場合は文全体を単位として談話構造解析を行う。

3.3.2 箇条書きや参照表現

箇条書きの文章を要約に残す時は，全項目を短くして残す。また，「これに基づいて」「以下の様に」などの参照表現がある時は，その参照先が要約に含まれるようにする。

5. 評価実験

5.1 解析事例

論文の解析事例を図 3 に示す。白い部分が要約文を表す。補足的な記述を省略できている。

固有表現とは，人名，地名，組織名などの事物に特有の表現をもつ単語を表すもので，主に，固有名詞が相当する。要約において，固有表現は，その特有の事物を表していることから，欠かせない情報であり，話題の中心になることが多い。よって，本研究において，重要語を選択するにあたり，固有表現の影響を及ぼすことができるように，固有表現得点として実現した。具体的には，SAGE が出力する文節域の referent が存在する場合，固有表現であるので固有表現得点として 50 点を加点する。

図 3 解析事例

5.2 学習

重要語の学習にはランキング SVM を用いた。学習用データとして，8000～10000 字程度の学会論文 7 件 5286 事例について，人手で作成した要約率 10%，30%，50%の正解データを用いた。

5.3 評価方法

精度評価は，被験者 6 名による主観評価で行った。評価データとして学習用データとは異なる論文 2 件について，要約率 30%で要約した結果と原文を比べ，「重要な内容がカバーできているか」を，全体を 3 つのエリアに分けそれぞれ 10 点満点で採点してもらった。また，日本語として意味が通じない文をチェックしてもらい，要約結果における割合を計算した。

5.4 実験結果

実験結果を表 1 に示す。結果は被験者による採点結果の平均値である。

表 1 評価実験結果

	内容カバー(10点満点)				日本語の間違い
	はじめに 概要	手法	実験 結論	平均	
論文A	7	6.33	10	7.78	9.3%
論文B	6	6.33	8.33	6.89	6.7%

6. 結論

章節構造に基づく自動要約手法を提案した。評価実験の結果，90%以上の精度で日本語として正しい文を出力し，全体的に内容もカバーした要約文を出力することができた。一方で，手法などの部分で重要な内容をカバーしきれていない部分もあった。得点の精緻化などの必要がある。

7. 参考文献

[1] 原田実, 水野高宏: “EDR を用いた日本語意味解析システム SAGE ”, 人工知能学会論文誌, Vol. 16, No. 1, pp. 85-93 , (2001).
[2] 田中信彰, 面来道彦, 野口貴, 矢後友和, 韓東力, 原田実: “意味解析を踏まえた自動要約システム ABISYS”, 言語処理学会論文誌, Vol. 13, No. 1, pp. 143-164, (2006).
[3] 梅澤俊之, 原田実: “センタリング理論と対象知識に基づく談話構造解析システム DIA”, 言語処理学会論文誌, Vol. 18, No. 1, (2011) (印刷中).
[4]SVM Light: http://svmlight.joachims.org/

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.