複数の情報を利用した意味段落分割に基づく文書理解支援

2Q－3
西澤信一郎
富士通株式会社

1. はじめに

計算機の発達に伴う電子化文書の増加がいわれて久しい．実際にCD-ROMにより配布される電子化文書やワードプロセッサにより作成・蓄積される文書などが身近な例としてすでに存在する．このような文書，特に規模の大きい文書を対象とした作業を計算機上で行なう場合，次のような欲求を感じることが多い．

・文書中の話題の概略をまず把握したい．
・自分が必要とする部分へ素早くたどり着きたい．
・ある話題に関連する部分をまとめて扱いたい．

これらに応える方法として，本論文では我々が現在考案中の，文書の意味段落分割に基づく文書理解支援機能を紹介する．また，実際に用いる意味段落分割の方法についても述べる．

2. 文書理解支援機能の概略

現在我々が考案中の文書理解支援機能の動作概略は次のようなものである．

(1)対象とする文書を解析し，意味段落構造を与える．解析時に用いる情報や手法については3節で述べる．
(2)各々の意味段落から抜粋を作成し，ガイドとして提供する．ガイドには現在対象とする意味段落からの抜粋に加え，類似した抜粋を持つ意味段落の情報を「類似段落」としてあわせて表示する．
(3)ガイドに表示する意味段落(抜粋文)と元の文書の間にリンクを保持し，互いの情報を参照できるようにする．

以上の動作によって作成される情報の利用の様子を図1.に示す．

ガイドとして表示される意味段落あるいは同時に表示される類似段落を選択し，図中の(a)のように元の文書を参照したり，元の文書からガイド情報を(b)のように参照し直す，などのように概略情報と詳細情報の間を自由に行き来することで話題の概略の把握や必要な箇所の素早い発見などを支援できると考えている．また，図中の(c)のように現在着目している意味段落に対する類似段落を前後に追跡する機能を用いることで，ある話題に関連する部分を集中的に扱うという作業を支援できると考えている．

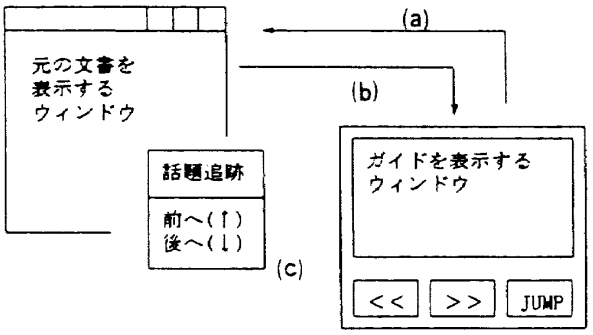

図1. 文書理解支援機能

3. 意味段落分割の手法

本論文で紹介する機能は，文書を意味段落へ分割することを処理の第1段階としている．本節では，現在我々が予定している意味段落分割の手法の概略と，そのなかでも主に検証を進めている，重要語の連鎖を利用した手法について説明する．

3.1. 概略および利用する情報

本論文で述べる機能では，意味段落分割の際に文書中の次の情報を用いる．

(1)ファイルの分割位置
(2)章・節などの形式段落
(3)文中の接続表現
(4)重要語(名詞)の連鎖情報

以上の情報について，文書中の該当箇所(形式段落の境界，話題転換の接続詞を含む文の直前，

A Concept of Document Understanding Support System based on Text Segmentation with Complex Information
Shin'ichiro Nishizawa
Fujitsu

ある重要語が連続して出現する範囲の先頭など)へ得点をつけ，それが別に定める閾値Th1を超える箇所を意味段落の境界とするという手法をとる．なお，(4)は3.2節で説明する．

3.2. 重要語の連鎖を利用した意味段落分割

文書中からの重要語の抽出およびその連鎖の利用方法は次のとおりである．

(1)文書全体を，それぞれが同じ数の文を含む仮の領域に分ける．

(2)各領域の内部に出現する全名詞について，その重要度をtf.idfにより次のように計算する．

領域Iでの名詞Jの重要度をW(I,J)とすると

W(I,J) = tf(I,J) × idf(J)

tf(I,J) = (領域Iでの名詞Jの出現回数)÷(文書全体での名詞Jの出現回数)

idf(J) = (文書全体での仮領域数)÷(文書全体での名詞Jの出現領域数)

(3)各領域毎に上記のW(I,J)が別に定める閾値Th2を超える名詞を選び，重要語集合Word(I)を作成する．また，文書全体でWord(I)をマージして重要語集合WORDを作成する[西澤97]．

(4)文書全体で，WORDおよびWord(I)を参照して3.1節で述べた得点付けを次の手順で行なう．

(a)着目する文が仮分割時に属した領域Xを決定する．

(b)WORDおよびWord(X)の双方に含まれる名詞の得点をP1，WORDのみに含まれる名詞の得点をP2とする．このとき，P1 > P2 とする．

(c)WORDに含まれる各名詞について，それが着目する文に含まれ，かつ直前の文に含まれないならばP1あるいはP2のいずれか適切な得点をその文に加える．

WORDに含まれる各名詞について，それが着目文に含まれず，かつ直前の文に含まれるのならばP1あるいはP2のいずれか適切な得点をその文に加える．

前者は重要語の連鎖の開始位置に相当し，後者は終了位置に相当する．

図2.に，上述した(1)～(3)の概略を示す．

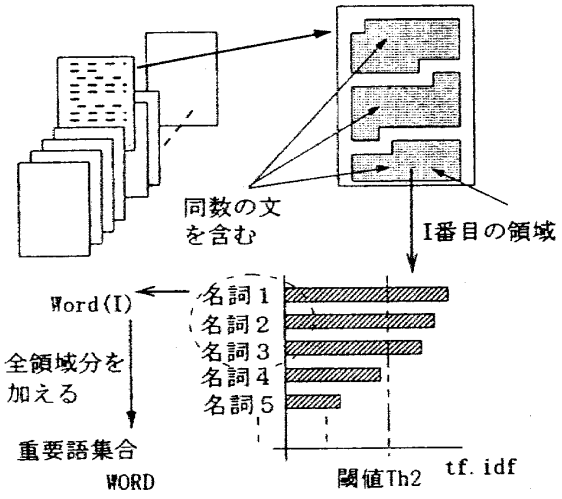

図2. 仮分割による重要語抽出

4. 現状と今後

現在，3.2節で述べた意味段落分割手法について，その有効性を検証するためのプロトタイプを作成している．ここでは，主に次の様なパラメタ値を検証の対象とする．

(1)仮分割の際に各領域に含まれる文の数

(2)区切決定のための点数の閾値Th1

(3)重要語抽出用の重みの閾値Th2

(4)区切決定のための点数P1，P2の比率

また，3.1節で述べた情報は文書の表層から得られる(意味解析を必要としない)ものである．同様の情報として，さらに文のタイプの変化による得点付けを行なう方法[田村98]などもある．プロトタイプによる検証の後には，3.1節で述べた情報の組み合わせ方や他に有効な情報があるか，などの検討も必要となると考えられる．

謝辞

本論文の件に関する議論に加わってくださった㈱富士通研究所メディア統合研究部の皆様，および貴重なご意見を下さった横浜国立大学の中川裕志教授，森辰則助教授に感謝いたします．

参考文献

[西澤97]西澤信一郎，森辰則，中川裕志．文書内の名詞の出現頻度を用いた段落分割．言語処理学会第三回年次大会 pp.389-392. 1997.

[田村98]田村直良，和田啓二．セグメントの分割と統合による文章の構造解析．自然言語処理 Vol.5 No.1 Jan. 1998.