文書への意味属性付与のための意味辞書の拡張

6F−7

中本　幸夫*　野上　謙一*　矢島　真人**　田野崎　康雄**
*:東芝ｺﾝﾋﾟｭｰﾀｴﾝｼﾞﾆｱﾘﾝｸﾞ(株)　**:(株)東芝

1．はじめに

ワープロ等の入力装置によって大量の文書が電子化されてくると、いかに一括管理し、要求に応じた文書を検索しユーザに提示するかが大きなポイントとなる。検索された文書が多いと、そのすべてをいちいち読んでいる余裕がないという現実的な問題を抱えている。このような現状では今後、文書の内容を把握してユーザの要求にもっとも合致する文書を提示できるシステムが望まれる。そのためには文書の論理構造を解析しただけでは、文書の内容を把握するには不十分であり、文書の流れが明確になるような意味的な文書構造を解析する必要がある。

そこで我々は、文章部分を意味的に解析し、『目的』、『課題』等の文書の内容を把握するための鍵となる意味属性を意味辞書を使って文章単位に付与する手法を開発した[1][2]。この意味辞書では構文レベルで同じ意味属性を持つものをまとめて構文パターンとして登録してある。今回は、『背景』や『意見』等の意味属性を加え意味辞書の拡張を行った。

本報告では、意味辞書の拡張内容と評価結果について述べる。

2．意味辞書

2−1．開発手順

文書の内容を把握するための鍵となる属性を文書の意味属性と呼ぶ。意味属性は「背景」「目的」「結果」「課題」などが考えられる。文書の章節等の特定部分にこれらの意味属性を付与することで、文書の大意などを簡潔に把握することができる。文書の特定部分に意味属性を付与する上で大切なのは、意味属性について明示的に述べている文である。このような文を内容明示文と呼ぶ。

今回拡張する意味辞書は、任意の文書から内容明示文を抽出するための構文規則を記述したものである。この意味辞書を開発する手順は、

（1）内容明示文の抽出
（2）内容明示文の特徴調査
（3）構文パターンの抽出

からなる。これらの作業は「東芝レビュー」の31文書（約1．5千文）を対象として行った。

2−2．内容明示文の抽出

まず、内容明示文の抽出は、3名が同じ技術論文を全文読み各人が人手によって行った。複数の人で作業を進めることにより内容明示文の抽出漏れを減らし各自の観点から幅広く抽出することを目指した。今回は技術文書でどの文書にも比較的よく含まれている『目的』『方法』『結果』『背景』『意見』『特徴』『課題』『内容紹介』の8種類の内容明示文に注目した。『内容紹介』とはこれから何を述べようとしているのかが分かるところで、例えば「ここでは〜について述べる。」のようなものである。また、特定の語句（「目的」「結果」など）や言い回しにとらわれることなく抽出した。

表2-2が抽出した内容明示文の数である。

表2-2　内容明示文数

『目的』	『方法』	『結果』	『背景』
50	130	88	118
『意見』	『特徴』	『課題』	『内容紹介』
85	265	49	230

全体文数に対して人手で抽出された内容明示文の割合は、27.2%であった。

2−3．内容明示文の特徴調査

内容明示文を抽出する際に内容明示文とする決め手となった語句や言い回し（ターゲットと呼ぶ）を指摘した。表2-3が『課題』『意見』『背景』のターゲットの例である。

表2-3　内容明示文のターゲット

意味属性	ターゲット
『課題』	〜が重要な課題である。 今後〜する予定である。 〜が今後とも重要である。
『意見』	〜と思われる。 〜することを期待している。 〜させたい。
『背景』	近年、〜となってきた。 いまや〜となってきている。 〜が強く望まれている。

これらの例からわかるようにどの内容明示文にも特徴的なターゲットがある。『課題』のターゲット「今後〜する予定である。」は見方を変えれば『意見』ととることもできる。複数の人によって内容明示文を抽出すると同じ文に異なる意味属性が付与される場合があり、『課題』と『意見』の組み合わせの文がもっとも多かった。『背景』は他の意味属性とは異なり、ターゲットが内容明示文の中に含まれておらず前後の文から判断される場合が多く、特徴づけに難しいものがあった。

Semantic Lexicon for Document Structure Method
Yukio NAKAMOTO, Ken'ichi NOGAMI, Makoto YAJIMA, Yasuo TANOSAKI
* Toshiba Computer Engineering Corp.
** Toshiba Corp.

２－４．構文パターンの抽出

先に抽出した内容明示文を構文解析し、指摘したターゲットを構文的な特徴のパターンとして取り出す。このパターンを登録したのが意味辞書である。構文パターンはターゲットを含む内容明示文から作成した。

構文パターンを作成した３１文書で意味辞書の検討を行った。構文パターンの作成と同一の文書で検討するので人手によって抽出した内容明示文はすべて抽出することができるが、内容明示文以外の文も抽出している。ここでは、意味辞書によって抽出された文のうち内容明示文である割合（抽出成功率）を調べた。表2-4が作成した構文パターン数と抽出成功率である。抽出成功率は意味属性別に69.9～97.7%である。

表2-4　構文パターン数と抽出成功率

意味属性	目的	方法	結果	背景
内容明示文数	37	121	71	86
構文パターン数	22	48	27	61
抽出成功率(%)	88.1	87.7	88.8	69.9
意味属性	意見	特徴	課題	内容紹介
内容明示文数	76	85	42	66
構文パターン数	44	52	35	43
抽出成功率(%)	89.4	88.5	97.7	97.1

(注)内容明示文数は構文解析成功数

３．評価実験

３－１．実験

意味辞書に登録した構文パターンが内容明示文の抽出に有効かどうかを評価するため、構文パターンを抽出した文書を除く「東芝レビュー」５４５文書（約３８千文）を用いて、次のような実験を行った。

（１）評価用データを構文解析し解析が成功した文を取り出す。

（２）（１）から意味辞書を用いて文抽出を行う。

（３）（２）が内容明示文であるかどうかを人手によって調べる。

（４）検証用データから無作為に選んだ１０文書について原文書を読んで内容明示文を抽出し、（３）でも抽出されているかを調べる。

３－２．結果

意味辞書の評価結果を表3-1, 3-2, 3-3に示す。

評価用データで構文解析が成功した35722文から意味辞書により6301文が抽出され、このうち5120文が内容明示文であった。抽出成功率は８１％であった。

次に、東芝レビュー１０文書によって、人手によって抽出した内容明示文が意味辞書によってもマッチする割合（カバー率）は７５％であった。

表3-1　評価用データ

評価データ文数	545文書　38098文
構文解析成功文数	３５７２２文

表3-2　意味辞書の成功率

	目的	方法	結果	背景
(a)意味辞書で抽出した文数	333	1323	1345	1051
(b)(a)のうちの内容明示文数	276	1174	1173	453
抽出成功率（%）(b)/(a)×100	８３	８９	８７	４３
	意見	特徴	課題	内容
(a)意味辞書で抽出した文数	938	685	250	446
(b)(a)のうちの内容明示文数	816	536	225	438
抽出成功率（%）(b)/(a)×100	８７	７８	９０	９８

表3-3　意味辞書のカバー率

検証データ	東芝レビュー　１０文書
(a)人手によって抽出した内容明示文数	３１３
(b)意味辞書により抽した内容明示文数	２３４
カバー率（%）(b)/(a)×100	７５

このように、『目的』『結果』『特徴』などの意味属性を付与できることが確認できた。構文解析を行い属性付けする、構文パターンを用いた方法により、今までより意味的な属性付与ができると考えられる。

しかし、『背景』の成功率が他の属性に比べて低いのは、この抽出方法が一文のみに注目しその文中に使われている語句や言い回しをトリガーにしているためであり、「背景」の成功率を上げるには文書の構造や前後の文も考慮する必要あると考えられる。

４．あとがき

文書の内容を簡潔に把握できるような内容明示文を抽出することができる意味辞書を開発した。意味辞書には、『背景』『特徴』などの意味属性を述べている内容明示文を抽出するために構文パターンを３３２個登録した。この意味辞書を用いて、評価用の文書から内容明示文の抽出実験を行ったところ抽出成功率８１％が得られ全文の約１割に属性付けができ、意味辞書の有効性を確認した。

参考文献

［１］岩井他：意味解析を用いた文書構造化手法，情報処理学会第43回全国大会論文集、1991

［２］矢島他：文書への意味属性付与のための意味辞書の開発、情報処理学会第43回全国大会論文集、1991