概要マニュアル文を対象とした構造解析の試み

3C-4

中茎 洋一郎*　村木 一至**
日本電気㈱ C&Cシステム研究所* C&C情報研究所**

1．はじめに

近年、文章の要約・理解を行うシステムの研究・開発が盛んに行われている。とくに、与えられた文章のどの部分に筆者の言いたいこと、強調したいことが書かれているかを知ることは重要な課題となっている。

従来のアプローチとしては、記述対象に関する知識を利用して解析を行うものがある［1，2］。しかし、実際には膨大な知識を必要とするため、実用的なシステムの実現は困難である。

もうひとつのアプローチは、最頻出語や接続語等に注目して重要な文を決定していくものである［3，5］。この場合、最頻出語等による判定では、重要文の選択が必ずしも正確に行われない場合も多いという問題点がある。

従って、現実的な方法で、より正確に文章の解析を行うメカニズムが要求されている。そこで本稿では、概要マニュアルの文章を対象として、用言の種類に着目し、強調文の抽出、段落分け等の文章構造の解析を進める方式を提案する。本方式では、記述対象に関する知識を用いずに解析を進めている点に特徴がある。実験システムを試作し、実際の概要マニュアル文に適用、評価を行った。

2．研究の枠組み

入力文章に対して各文毎に形態素解析、構文解析を行い、概念間の関係を示すｔｒｅｅ構造で表される中間表現［4］に変換する。

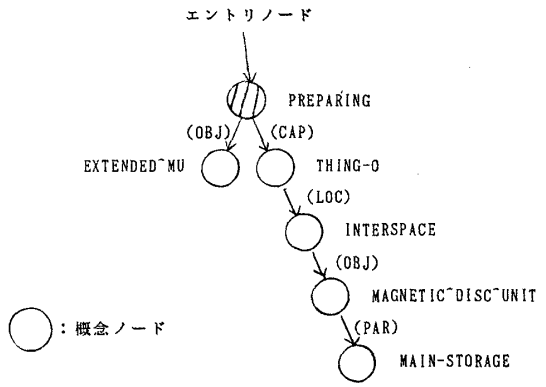

図1　中間表現の例

An approach for analyzing manuals.
Yoichiro NAKAKUKI, Kazunori MURAKI
NEC Corporation

通常の文は、その中心となる用言に対応する概念がｔｒｅｅのルートノード（エントリノード）となる。例えば、「主記憶装置と磁気ディスク装置の中間に位置するものとして拡張記憶装置を用意しています。」という文の中間表現は図1のようになる。

各概念ノードに対しては、さらにさまざまな属性情報が付けられ、元の文の内容が表現されている。与えられた文章の各文に対する中間表現を順に並べたものを入力としている。

概要マニュアルは、製品やシステムの仕様、特長について述べるためのものであり、機能・性能の詳細、その必要性、実現方法について述べた文が多数を占めている。従って、使われている用言に注目することで、その文の重要度、役割りについて推定できる場合が多い。このような特徴を利用して文章構造の解析を試みることにする。

3．文章解析法

概要マニュアルの文の中から、特に強調したい内容を述べた文（強調文）の抽出を行い、内容的にひとまとまりの段落の決定を行う方法について以下に示す。

3．1強調文の抽出方法

概要マニュアルでは、主として製品やシステムの特長、何が可能か、どのようにそれを実現しているのか等が強調したい点となる。従って、中心となる用言に着目することでかなり正確にそのような強調文を見分けることが可能である。そこで、用言が次のような意味、状態を表している文を強調文と判断することにする。

- 可能　（……することができます、等）
- 過去　（……を実現しました、等）
- 現在継続中の状態（……を高めています、等）
- 動作の結果として存在する状態（……を採用しています、等）

一方、上記以外の文は説明文として扱われる。

（例）　……が重要です

3．2　段落分けの方法

入力文章を章・節の番号やその他の見出し等によって、数文から十数文のひとまとまりの文章に分割する。その一連の文章の内部構造を明かにしていくため、さらにその中に含まれる各強調文を中心とした段落に分割していく処理を行う。

●高度化する科学技術分野では，大規模シミュレーションへのニーズが高まっています。

●このシミュレーションの多くは偏微分方程式で表わされ，さらには連立一次方程式や固有値問題を解くベクトル処理に帰着されます。

◎ＳＸシステムでは，ＮＥＣ日本電気の総力をあげてベクトル処理における“Ｇ”の領域にチャレンジし，最高１３００ＭＦＬＯＰＳという超高速処理を実現しました。

図２　概要マニュアル文の段落の例

概要マニュアルの意味段落の構造として最も多いパタンは、図２に示すような、いくつかの説明文に引き続いて強調文が記述されるというものである。

この例では、シミュレーションやベクトル演算の必要性について説明する２つの文章（説明文●）に続いて「‥‥を実現しました」という最も強調したい内容が記述されており（強調文◎）、全体として、ひとまとまりの段落となっている。

そこで、段落分けの方法として、基本的にはいくつかの説明文とそれに引き続く強調文とで１つの段落としている。ただし、処理の対象となるひとまとまりの文章が強調文で終っていない場合には最後に現われる強調文以降の説明文はその強調文と同じ段落とする。段落分けの例を図３に示す。

文１●
文２●
文３◎

文４◎

文５●
文６◎

文７●
文８◎
文９●

●：説明文
◎：強調文
---------- ：段落の境界

図３　段落分けの方法

さらに、次に示すようなキーとなる接続詞・指示語に注目してより正確な段落分けを行っている。

－ 強調文で「また」、「さらに」が用いられている場合、その文から新しい段落が始まるものとする。

－ 強調文Ａに続く文Ｂにおいて、文Ａの内容を「この」によって参照している場合、文Ｂを文Ａと同じ段落とする。

４．実験結果

提案する方法にもとづき、システムを試作し、実際の概要マニュアルの一部（約２００文）を入力として評価を行った。

強調文と考えられる文は全部で６４文あり、そのうち５７文が同システムで強調文と判定された。残りの７文は「‥‥の機能があります。」、「図××に‥‥を示します。」といった形のもので、用言からだけでは判断できないものであった。

逆に、強調文ではないものが強調文であると判定されたものが６文あった。これらは、たとえば、「‥‥のニーズが高まっている。」、「‥‥はＡとＢに分けられる」といった形で、一般的な事柄に関する記述文でありながら、主語や動作主等の推定を行っていないために強調文と判断されたものと考えられる。

一方、段落分けの結果に関しては、極端に不自然な部分はなかった。ただし「この」等の指示語が指す内容の推定を行っていないため、適切な段落分けができていない部分も見受けられた。

文章の種類が概要マニュアル文という以外、記述対象については特に限定せずに、かなり良好な解析結果が得られたといえる。今後はさらに精度の高い解析を実現していくことが必要であると思われる。

５．まとめと考察

本文では、概要マニュアルを対象とした文章解析方法を提案した。概要マニュアル文の特徴を用いることで強調文の検出、段落分けの処理をかなり実用的な正確さで行うことが可能となった。

このように文章の種類を限定して、その構造解析を行う手法は概要マニュアル文以外にも応用することが可能であると考えられる。

また、より正確な解析を行うためには、重文や複文において、それを構成する各単文に対する処理を付け加えることや、中間表現を作成する際の形態素／構文解析の段階での解析情報を利用することなどが今後の課題である。

【参考文献】

[１] DeJong, G., "Skimming newspaper stories by computer", IJCAI '77

[２] Fum, D. et al., "Evaluating importance: a step towards text summarization", IJCAI '87

[３] 鈴木、他，"科学技術文献の要約システムについて(1)(2)" ３３回情処全大(1986)

[４] 村木，"知識ベースと，言語に独立な中間表現とを用いた日英機械翻訳システム" 日経エレクトロニクス 12.17 (1984)

[５] 喜多，"説明文を要約するシステム" 情処学会自然言語処理研究会 63-6(1987)