文章構造解析システムにおける同語反復解析処理

田中　理[†]　柴田昌宏[†]　福本淳一[††]

[†]（株）沖テクノシステムズ ラボラトリ　　[††] 沖電気工業（株）

1　はじめに

論説などの文章においては、ある事がらに関する書き手の主張が読み手に伝わるように論旨が展開される。現在、我々は、この論旨の展開構造を文章の構造であるととらえ、論述内容を把握することを目的とした文章構造解析システムの開発を行なっている[1]。

この構造解析のために必要な情報として、書き手が明示的に展開を示す手段である接続詞のほか、文末表現などの表層情報、および同語反復などの照応関係があり、これらを用いることによって文間関係の解析を行なうことができる[1][2]。

本稿では、文章構造の解析のため、同語反復に焦点をあて、これが文間でどのような役割を持っているかを分析する。そして、これが文間関係にどうかかわっているかについて述べる。

2　同語反復現象

文章構造解析システムの全体構成を図1に示す。

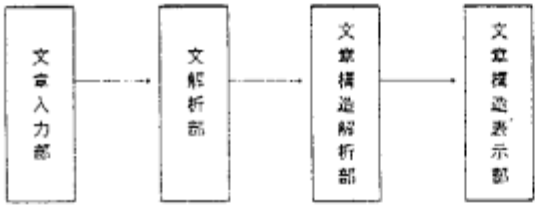

図1　システム構成図

このシステムは、文章ファイルを入力として、同語反復解析処理などにより文の持つ情報を取り出し、それらを用いて文章構造の解析を行ない、その結果を表示するものである。

一般に、文章においてはいくつかの話題が展開されており、話題が転換されている部分で区切ることにより、文章をいくつかの内容的なまとまりに分割（以下、段落分けと呼ぶ。）でき、文章構造の抽出が可能であると考えられる。

文章構造解析のために必要な情報として、接続詞や文末表現などの表層情報があり、これらを用いることによって文間関係の解析を行なうことができる。例えば、文頭に接続詞のある文があり、接続詞が逆接の意味であれば、その文と直前の文との文間関係は“逆接”であることがわかる。また、文末が「…か。」という問い掛けになっていれば、その文と次の文との文間関係は“呼応”であることがわかる。ただし、これらの情報だけで解析できるのは隣合った二文間のみであり、文章をいくつかの段落に分けることはできない。

文章をいくつかの段落に分けるためには、話題が連続している部分を取り出すことが必要である。そのため、同一の話題について述べるときには、同じ名詞または同じ言い回しを、隣合う文だけでなく複数の文に渡って使用する現象（以下、同語反復現象と呼ぶ。）があることに着目して、その調査を行なった。

3　同語反復の分類

以下に、昭和62年10月30日付朝日新聞朝刊社説記事に見られる同語反復の現象例を示す。（番号は、社説先頭からの文番号を示す。アンダーラインは、同語に対する修飾語を示す。）

a.　31.　労働条件を明記した雇い入れ通知書を普及させるべきだし、…。
　　34.　現状では、パートの労働条件は企業によってまちまちだ。

b.　1.　働く女性がふえ続け、労働力人口の4割を占めるまでになった。
　　2.　雇われる女性も急増した。

上の現象例を見ると、a.では、同語反復の前の語が「労働条件」、後の語が「パートの労働条件」となっていて、後の語は前の語の一部を指すと思われる。すなわち、後文は前文を“補足”していると考えられる。

また、b.では、同語反復の前の語が「働く女性」、後の語が「雇われる女性」となっていて、後の語は前の語を言い換えていると思われる。すなわち、後文は前文を“反復”していると考えられる。

このことから、同語反復はそれを修飾する語の有無、および各々の同語に対する修飾語によって、文間関係も異なると考えられる。

そこで、修飾語の有無に着目して、同語反復を次の4タイプに分類した。ここで、先行詞とは、同語反復における、前の同語のことを、照応語とは後の同語のことをいう。

A. 修飾なし：修飾するものがない
B. 修飾_先：先行詞のみ修飾するものがある
C. 修飾_照：照応語のみ修飾するものがある
D. 修飾_異：修飾するものが各々異なる

4　調査方法

昭和62年10月の朝日新聞社説記事5編を用いて、調査を行なった。

調査対象とする同語反復は、以下の検索アルゴリズムにより決定した。このとき、同一文中に同語反復が含まれていても、最初の同語のみを同語反復の対象とした。

(1) 文章の先頭の文を検索対象文とする。
(2) 検索対象文の先頭から順に、品詞が名詞である語を検索し、その語を検索語とする。
(3) 検索語と品詞および表記が同一である語（以下、同一語と呼ぶ。）をその後文から検索する。

Repetitive Words Analysis in Text Structure
[†]Osamu TANAKA, Masahiro SHIBATA
[††]Jun-ichi FUKUMOTO
[†]Oki Technosystems Laboratory, Inc
[††]Oki Electric Industry Co., Ltd
*本研究は、第5世代コンピュータプロジェクトの一環としてICOTからの委託で行われたものである。

(4) 検索した結果、

(i) 同一語が存在しない場合

(a) 検索対象文が文章の最後の文のとき、処理を終了する。

(b) 検索対象文が文章の最後の文以外のとき、検索対象文の次の文を検索対象文として、(2) 以降の検索を行なう。

(ii) 同一語が存在する場合

(a) 検索語を先行詞、同一語を照応語とする。

(b) 次に、検索語に修飾する語を加えて、(3) 以降の検索を同一語がみつからなくなるまで行ない、そのときの先行詞、および照応語を同語反復のデータとする。

上のアルゴリズムで求めた同語反復について、以下の2点に着目し、評価および調査を行なった。

1. 同語反復のタイプ毎の文間関係の割合

同語反復のタイプにより、次の文間関係が決定できるかどうかについて評価した。

a. 補足　前文の内容を補足する内容を後文に述べる。
b. 転換　前文の内容から転じて、別個の内容を後文に述べる。
c. 累加　前文の内容に付け加わる内容を後文に述べる。
d. 反復　前文の内容と同等とみなされる内容を後文に重ねて述べる。
e. その他　文間関係を決定しにくい場合。

2. 同語反復のときの文間関係と文間距離の関係

文間の距離がどのくらいのときに同語反復は多くみられるのか、また、文間の距離が離れていても同語反復により段落分けができるかについて調査した。

5 調査結果

1. 同語反復のタイプ毎の文間関係の割合

表1に、タイプ毎の文間関係の割合を示す。

その結果、タイプBのときには文間関係が"転換"になるものが約54%、タイプCのときには文間関係が"補足"になるものが約41%あった。

表1 同語反復のタイプ毎の文間関係の割合（単位%）

タイプ	補足	転換	累加	反復	その他	計	データ数
A	4.3	31.9	34.0	12.8	17.0	100.0	47
B	1.9	53.7	18.5	7.4	18.5	100.0	54
C	40.8	20.4	12.2	8.2	18.4	100.0	49
D	3.4	36.7	33.3	13.3	13.3	100.0	60
						計	210

2. 同語反復のときの文間関係と文間距離の関係

表2に、文間の距離毎の文間関係のデータ数を示す。

その結果、同語反復は、文間の距離が1の場合が最も多く、距離が離れる程減少していく傾向があった。

文間関係が"補足"になるのは、文間距離が4以下のときだけであった。

最も多かった文間関係は"転換"であり、約36%あった。

表2 同語反復のときの文間関係と文間距離の関係

距離	補足	転換	累加	反復	その他	計
1	10	16	15	8	4	53
2	6	5	12	4	5	32
3	5	5	5	1		16
4	4	11			2	17
5		4				4
6			5	2	3	10
7		1	3	1	2	7
8～		34	12	6	19	71
計	25	76	52	22	35	210

6 考察

調査結果をもとに、以下の考察を行なった。

1. 同語反復のタイプと文間関係

タイプBのときは文間関係が"転換"になることが多く、タイプCのときは文間関係が"補足"になることが多いことから、同語反復のタイプにより文間関係を決定するには、タイプBおよびタイプCが有効であると思われる。

文間関係が"補足"のときは、先行詞を含む文から照応語を含む文までは一つの話題についての連続性があると考えられ、それらを一つの段落とみることができる。

また、"転換"のときは、先行詞を含む文から照応語の直前の文までは話題が連続していて、照応語を含む文で話題が変わっていると考えられ、それらを一つの段落とみることができる。

2. 同語反復のときの文間関係と文間距離

同語反復のときの文間距離が、4以下のとき"補足"になりやすい傾向があるが、全体的にはバラツキがあり、必ずしも文間距離により、文間関係が決定できるわけではない。

7 おわりに

同語反復のときの文間関係と文間距離には、特に関係がないことがわかった。また、同語反復のタイプによっては段落分けができる場合があることを示した。

その際、同語反復を含む文は、必ずしも隣合っていなくてもよく、他の情報（接続詞、文末表現などの表層情報など）で決定できないときでも有効であると考えられる。

ただし、同じことをいうのに異なった表現を用いる場合もあり（例えば、「女性」と「婦人」など）、同語反復以外にも段落分けに有効な情報があると考えられる。

今後は、5社説のみでなくもっと多くの社説を調査して、今回の成果を確認評価していく予定である。

また、同語反復のタイプによって段落分けのできない部分については接続詞などの表層情報の利用法を検討し、文章構造の解析精度を向上していく予定である。

【参考文献】

[1] 福本淳一：筆者の主張に基づく日本語文章の構造化, 情報処理学会自然言語処理研究会報告, 78-15, (1990).

[2] 斉藤, 柴田, 福本：文章構造化のための文の連接関係の解析, 情報処理学会第43回全国大会, 5H-4, (1991).