結束チャートを用いた日本語文章の語彙的結束構造の解析

7B−5

佐々木 一朗*　増山 繁*　内藤 昭三**

*豊橋技術科学大学　**NTT基礎研究所

1 はじめに

自然言語処理における日本語文章の談話構造の解析は現在までに数多く試みられている。本稿では、談話構造を解析するための手法の一つとして、結束チャートと呼ぶデータ構造を新たに提唱し、それを用いて実際の文章の結束性を解析する。結束チャートとは、文章の各段落内及び、段落を跨っている、シソーラスに基づく意味分類のつながりをチャート形式で図示したものである。これを用いれば、段落内、又は、段落間に跨る意味分類が容易に把握できるので、文章全体の語彙的結束性に関する大域的構造の解明に利用できる。結束チャートの有効性を確認するために、実際の文章に対して結束チャートを構成し、これを利用したテキストの語彙的結束性の考察を行なった。テキストからのシソーラス中の語の抽出には、計算機を利用し、多義語の意味分類の曖昧性解消は、手作業により行なった。結束チャートの応用としては、キーワードの抽出や付与、文章のタイプ分類、かな漢字変換における同音異義語選択などが考えられる。

2 結束チャート

今回、シソーラスには、角川書店「類語新辞典」[1]（収録語彙数57145語）を用いた。この辞典では各語に3桁の分類番号が付されていて、それぞれ、大、中、小の各分類に対応している。今回は、この中の小、中分類に着目した。また、チャートが過度に繁雑にならないようにするために、本報告中の結束チャートには、どの中分類が、どの段落に出現しているかを図示した。結束チャートの作成は、以下の手順に従った。テキストからシソーラス中の語を抽出した後、段落毎に、各小分類、中分類の出現を調べた。今回、段落毎に意味分類の出現を調べた理由は、1) 1文中では、分類を調べるにはあまりにも語彙数が少ない(特に、天声人語でこの傾向が顕著)、2) 文章全体の語彙的結束性に関する大域的な構造を把握するためには、文を単位とするよりも段落を単位とする方が適切であると判断した、ためである。

今回、分析対象としては、天声人語10編と日経サイエンスの記事5編を使用した(段落数は天声人語で、総段落数60、平均段落数6、日経サイエンスで総段落数46、平均段落数9程度である)。天声人語と日経サイエンスを選んだ理由は、1) いずれも、充分練られた文章である、2) 扱われている分野や、段落中の平均文数の異なる文章によって、どのような傾向の違いが現れるかを比較する、ためである。

3 分析結果

今回の分析(表1～8)、及び、結束チャート(図1、図2)の分析から、次の事が明らかとなった。

1. タイトルに関連する意味分類の出現回数が非常に多いことが確かめられた。

2. 段落が短い場合、文章全体での中分類番号のまとまりが悪い。これは段落毎で、話題が変化していることを表していると考えられる。特に天声人語ではこの傾向が顕著である。

3. 日経サイエンスのように、限定された話題のみを扱っている記事においては、話題に関するある特定の語の出現頻度が非常に高くなる。つまり限られた特定の小、中分類項目の出現回数が多くなる。これは、後述の主題の認定、キーワード抽出等に有用な性質である。

4. 大分類で、分類番号100番台は、非常に出現回数が多い。これは指示、代名詞、数詞なども含まれているためである。

5. 隣接した段落間では、同じ中分類項目が継続して出現することが多い。またタイトルに関連する中分類は、最初から最後まで連続して出現することが多い。

6. 天声人語と日経サイエンスを比べると、天声人語は段落長が短いため、小、中分類の出現が持続しない。日経サイエンスでは出現する小分類、中分類の種類が天声人語と比べると多いにも関わらず、ひとたびある分類が出現すると連続して、複数の段落に跨り出現することが多く、特に図1の結束チャートからもわかるように、いくつもの段落を跨ぐような中分類の数が非常に多い。

4 考察

それぞれの文章で最初の方の段落にのみ出現する分類があることがわかった。この現象は、特に日経サイエンスの方で顕著に見られた。この分類が、その文章全体における何かキーとなる分類であるのか、また形式的な起承転結の"起"の部分に対応する分類であるのか、それとも単なるその段落内だけの話題なのか、今後更に検討して行きたい。

各段落内の出現回数別の小分類数の調査により、出現回数が1回のもののほとんどは、話題に対する雑音とみなすことができる。また、中分類については、隣接した段落間でのつながりが一番多く、またサイエンスのような絞られたテーマで書かれている文章ほど段落を跨っている分類数が大きくなる、つまりある分類(話題)が続いていて、それまでの話題が段落毎に大きく変わらないことが明らかになった。

また多義語に関して次のような傾向が見られた。

1. 多義語（つまり、2つ以上の意味分類を持っている語）の場合、一つの文章中では、一つの意味分類のみで使われる場合が多い。たとえば、衰弱という単語には「発病」と「盛衰」の2つの意味分類があるが、日経サイエンスのある文章(ガン治療についての話題)では、すべての出現が「発病」の意味で使われていた。またこの現象は、出現頻度の高い多義語について顕著に見られた。

2. 多義語においては、人間が判断して正しい意味が属する中分類の方が、他の語においてもその中分類に属する意味を持つことが多い。つまり、中分類でのまとまりが見られるという傾向にある。これは、多義語は、近傍(同一段落など)に出現する中分類と同じ分類を持つ意味で使われることが多いためと思われる。

多義語に関してのこれらの結果は、今後結束チャートの自動生成を試みる際に有用である。

5 むすび

本稿では、結束チャートを利用して、テキストの語彙的結束性を分析したが、この他にも、自然言語処理の分野で次のような結束チャートの利用が考えられる。

1. 文章全体の分類カテゴリーの分布を数量化し、そのパターンにより文章のタイプ分類を行なう。

2. キーワードの抽出や付与に応用する。

3. かな漢字変換における同音異義語選択へ応用する。

今後は、前章でも考察した、結束チャートの自動生成で必要となる多義語の処理方法を検討するとともに、ここにあげた結束チャートの利用法の検討を進める。

An Analysis of Lexical Cohesion in Japanese Sentences using the Cohesion Chart
Ichiro SASAKI*, Shigeru MASUYAMA*, Shozo NAITO**
*Toyohashi Univ. of Tech. , **NTT

謝辞

「類語新辞典」を計算機可読辞書の形で提供していただき、その使用許可を頂いた(株)角川書店に深謝する。

調査結果

	一段落内の出現回数	小分類の数	全小分類数に対する割合
各段落内の出現回数別の小分類数	1	707	72.3
	2	152	15.6
	3	53	5.4
	4	14	1.4
	5以上	52	5.3

表1．各段落内の小分類の出現回数(天声人語10編)

	段落数	個数
小分類が跨っている段落数	1	942
	2	88
	3	53
	4	6
	5	2
	それ以上	1

表2．小分類が跨っている段落数(天声人語10編)

	一段落内の中分類の出現回数	中分類の数	全中分類数に対する割合
各段落内の出現回数別の中分類数	1	255	48.1
	2	156	29.4
	3	73	13.8
	4	34	6.4
	5	8	1.5
	それ以上	4	0.8

表3．各段落内の中分類の出現回数(天声人語10編)

	段落数	個数
中分類が跨っている段落数(広がり)	1	512
	2	117
	3	49
	4	20
	5	5
	それ以上	6

表4．中分類が跨っている段落数(天声人語10編)

	一段落内の出現回数	小分類の数	全小分類数に対する割合
各段落内の出現回数別の小分類数	1	2677	50.4
	2	1031	19.4
	3	542	10.2
	4	329	6.2
	5以上	735	13.8

表5．各段落内の小分類の出現回数(日経サイエンス5編)

	段落数	個数
小分類が跨っている段落数	1	1939
	2	475
	3	236
	4	86
	5	66
	6	59
	それ以上	79

表6．小分類が跨っている段落数(日経サイエンス5編)

	一段落内の中分類の出現回数	中分類の数	全中分類数に対する割合
各段落内の出現回数別の中分類数	1	886	39.6
	2	505	22.6
	3	355	15.9
	4	224	10.0
	5以上	267	11.9

表7．各段落内の出現回数別の中分類数(日経サイエンス5編)

	段落数	個数
中分類が跨っている段落数(広がり)	1	310
	2	112
	3	69
	4	24
	5	40
	6	33
	それ以上	101

表8．中分類が跨っている段落数(日経サイエンス5編)

参考文献

[1] 大野晋、浜西正人 ： 類語新辞典　角川書店　１９８１

[2] 池上嘉彦 ： テクストとテクストの構造, 国立国語研究所(編)
談話の研究と教育1,pp.7－42, 大蔵省印刷局（１９８３）.

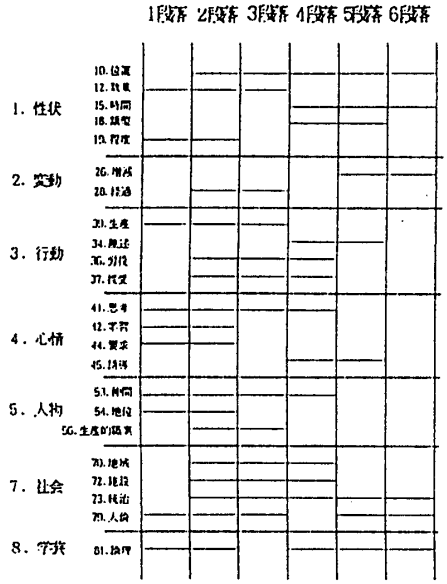

図1 結束チャート（天声人語）

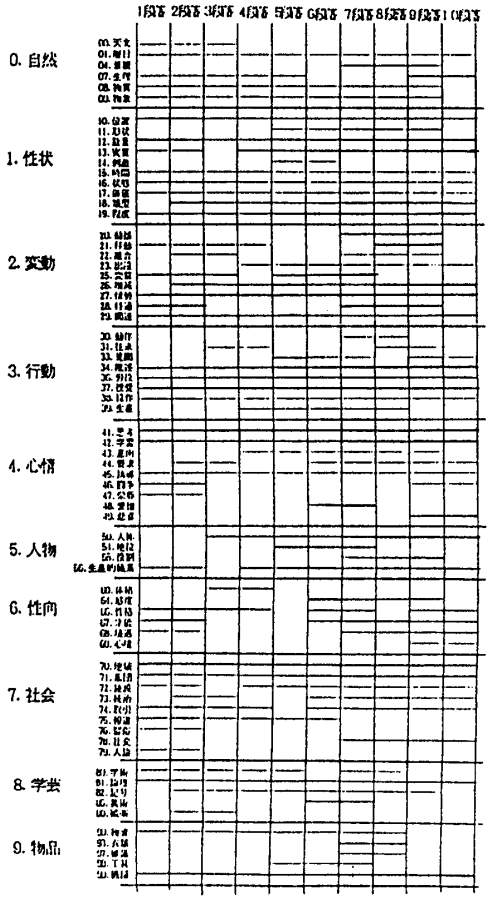

図2 結束チャート（サイエンス）