自然言語処理　95－8
(1993. 5. 20)

結束チャートの自動生成と日本語文章の語彙的結束構造解析への応用

佐々木 一朗*　　増山 繁*　　内藤 昭三**
{sasaki@toki. , masuyama@}tutkie.tut.ac.jp*　naito@atom.ntt.jp**

* 豊橋技術科学大学知識情報工学系
** ＮＴＴ基礎研究所
* 〒441, 愛知県豊橋市天伯町字雲雀ヶ丘1-1
** 〒180, 東京都武蔵野市緑町3-9-11

概　要

本稿では、日本語文章の談話構造の解析手法の一つとして結束チャートを提唱し、その構築方法及び自動生成法、利用法、応用などについて考察する。結束チャートとは、文章の各段落内及び、段落を跨っている意味分類のつながりを、シソーラスに基づいて解析し、その結果をチャート形式で示したものである。実際に、日経サイエンス及び天声人語の記事に対して結束チャートの自動生成を試み、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。

和文キーワード　結束チャート, シソーラス, 談話構造, 語彙的結束性, 文脈解析

Automatic Production of Cohesion Chart and its Application to the Analysis of Lexical Cohesion in Japanese Sentences

Ichiro SASAKI*　Shigeru MASUYAMA*　Shozo NAITO**

*Dept. of Knowledge-based Info. Eng., Toyohashi Univ. of Tech.
**NTT Basic Research Laboratories
*1-1, Azahibarigaoka, Tenpaku-cho, Toyohashi-shi, Aichi, 441, Japan
**3-9-11, Midori-cho, Musashino-shi, Tokyo, 180, Japan

Abstract

In this paper, we advocate a method of analyzing discourse structure of Japanese sentences using the cohesion chart, its automatic production and its application. The cohesion chart is constructed using the thesaurus of classifying words into large, middle and small categories and is used to indicate how words belonging to each middle category continue to appear among paragraphs in Japanese sentences. We try to automatically produce the cohesion chart for each article or essay, and discuss its application to the analysis of lexical cohesion in Japanese sentences.

英文 key words　Cohesion Chart, thesaurus, Discourse Structure, Lexical Cohesion, Context Analysis

1 はじめに

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提唱し、それを用いて実際の文章の結束性を解析し、あわせて、この結束チャートの自動生成を試み、その有用性を検討する。自然言語処理における日本語文章の談話構造の解析は、テキストの大域構造に注目する方法、局所的な接続関係に注目する方法など、現在までに数多く試みられている [4, 5, 6, 7]。本稿で提唱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分類(番号)毎に図示したものである。これを用いることにより、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。本稿では、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作成し、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、多義語選択については、手作業による方法と、自動選択による両方法を試みた。後者については、3種類の多義語選択の方法を試み、これら方法を比較した。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、文章の要約などが考えられる。なお、本稿の部分的結果を [1] で報告した。

2 結束チャート

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。この結束チャートというデータ構造を用いることにより、段落内、又は、段落間に跨る意味分類を容易に把握することができ、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。

また、結束チャートの自動生成を試みたが、その目標は、人間が作成した場合と同様の結束チャートを、早く大量に作り、処理を行なおうというものである。また、この結果をもちいて、文章の分類、タイプ分け等を計算機で的確に行なおうというものである。

3 シソーラス

シソーラスは、角川書店「角川類語新辞典」[2] (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、語彙的結束性の調査ではこの中の小、中分類に着目し、また結束チャートの生成では主に中分類のみに着目して使用した。これらの理由は、大分類では分類が粗過ぎて、語彙的結束性の調査及び結束チャートの構成に、あまり意味を持たなくなってしまい、一方、小分類を用いると、チャートが過度に繁雑になってしまうからである。そこで、本報告中の結束チャートでは、どの中分類が、どの段落に出現しているかを図示している。

4 使用テキスト

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。

5 シソーラス語抽出

シソーラス語抽出のために簡単な形態素解析(最長一致法による)プログラムをC言語で作成し、それによりシソーラス語を抽出した。この段階では、多義語に対しては、全ての意味分類を抽出する。

6 多義語選択及びチャートの作成法

結束チャート自動作成上の問題点は、上でも触れたテキスト中の5～9%を占める多義語(つまり、2つ以上の意味分類を持っている語)の処理である。そこで、まず手作業により多義語選択を行ない、結束チャートを作成し、テキストの分析を行ない、その結果を考察することにより、多義語選択における傾向を検討した。次に、この考察結果に基づく多義語の自動選択法を組み込み、結束チャートの自動生成を行なった。多義語選択を手作業によった結束チャートの生成については次

節で、多義語の自動選択による結束チャートの自動生成については後章で示す。意味分類の出現は段落毎に調べている。その理由は、以下の通りである。

1. 1文中では、分類を調べるには対象となる語彙数が非常に少ない(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。

6.1 多義語の手作業選択による結束チャートの作成

計算機でテキストからシソーラス中の語を抽出した後、手作業により段落毎に、各小分類、中分類の出現を調査した。その際の多義語の選択は、人間が意味を調べ、適切な分類を選択した。この結果をデータとして、後述の結束チャートの自動生成に用いたチャート描画プログラムによりチャートを作成する。また、今回作成したチャートには、段落間の語彙的結束性を見るために2つ以上の段落に跨っている分類のみを表示している。

7 語彙的結束性の調査結果

以下に調査結果を示す。

	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)

小、中分類の各分析(表1～8)、及び、結束チャート(図1他)の分析から、次のような特徴が明らかとなった。

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。

2. 段落が短い文章の場合、文章全体での中分類番号のまとまりが悪い。これは段落毎に、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。これは指示語、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い。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。

7.1 分類調査に対する考察

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であった。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、今後更に検討していきたい。

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。

8 結束チャートの自動生成

前章までの調査から得られた結果とその考察に基づき、結束チャートの自動生成を試みた。以下にその方法、結果、考察を示す。なお、プログラムは全てC言語を用いて作成した。

8.1 多義語の自動選択による結束チャートの生成

テキストからシソーラス中の語、及び分類を抜き出し、手作業で多義語選択を行ない作成した結束チャートの分析により得られた知見に基づいて多義語の自動選択プログラムを作成した。多義語の自動選択の方法は次の3種類を試みた。

方法1 多義語の意味分類の選択を行なわないで、全ての分類を使用する。

方法2 段落頭から現在処理を行なっている位置までの中分類の累積頻度(段落内頻度)が最大の分類を選択する。ただし、初めて出現する分類などの理由により、タイが起こる場合は、若い番号の分類を選ぶ。

方法3 テキスト頭から現在処理を行なっているところまでの中分類の累積頻度(総合頻度)が最大の中分類を選択する。この場合のタイブレイクは方法2の場合と同様である

方法1は実際には多義語の選択を行なわないものであるが、他の方法との比較のために用いた。方法2と方法3は、手作業による多義語で得られた結束チャートの分析及び、分類調査より得られた次のような観察に基づいている。

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. 最初に選択したのと同じ意味分類が、後の語においても選択されることが多い。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。

チャートの作成は、多義語選択の結果を一度ファイルへ落し、チャート描画プログラムによりチャートを描かせる方法をとった。図示は、X11R5上で、Xlibを使用して行なった(図1～4)。また、段落間の語彙的結束性を見るために、隣接した段落間で同時に出現している分類のみ(隣あった段落で続けて出現している分類)を扱うこととした。これにより、チャートが繁雑となるのを防ぐことができ、段落間の語彙的結束性の判断が容易になる。

9 手作業により作成したチャートと自動生成によるチャートの比較

以下の方法により、手作業により作成したチャートと、自動生成により作成したチャートを比較した。

1. 手作業によるチャートでは、多義語が適切に選択されていると仮定して、これを基準に自動生成による各方法との比較を行なう。
2. 比較は、パターンマッチングで行なう。
3. 分類がどの程度の段落に跨って出現しているかを見るために、2つ以上の段落間に跨っているもののみ(一つの段落内のみに出現しているものについては前述のように繁雑さを防ぐため除いた)を対象としてマッチングを行う。
4. 手作業によるチャートに出現している中分類が自動生成のチャートにも出現していれば正解として数え、手作業による結束チャートでの全出現分類数で割り、それを再現率として%で表す。
5. 手作業によるチャートに出現している中分類が自動生成のチャートにも出現していれば正解として数え、自動生成によるチャートでの全出現分類数で割り、それを適合率として%で表す。

以下にチャート自動生成の再現率結果を示す。

sample text	再現率 (%)		
	方法1	方法2	方法3
no1	94.9	76.9	89.7
no2	91.7	52.8	93.1
no3	86.9	63.9	85.2
no4	97.6	57.1	95.2
no5	86.9	55.7	90.2
no6	79.2	56.3	75.0
no7	78.8	61.5	94.2
no8	91.9	64.5	90.3
no9	82.5	59.6	93.0
no10	71.7	58.7	71.7

表9. 天声人語の再現率

sample text	再現率 (%)		
	方法1	方法2	方法3
no1	94.9	85.9	93.6
no2	93.2	85.7	95.9
no3	98.3	88.4	97.4
no4	98.2	86.0	98.5
no5	97.6	88.5	97.0

表10. 日経サイエンスの再現率

以下にチャート自動生成の適合率結果を示す。

sample text	適合率 (%)		
	方法1	方法2	方法3
no1	94.6	70.0	88.6
no2	90.9	10.5	92.5
no3	84.9	43.6	82.7
no4	97.6	25.0	95.0
no5	84.9	20.6	89.1
no6	73.7	22.2	66.7
no7	73.2	37.5	93.9
no8	91.2	45.0	89.3
no9	78.7	32.4	92.5
no10	60.6	29.6	60.6

表11. 天声人語の適合率

sample text	適合率 (%)		
	方法 1	方法 2	方法 3
no1	94.8	84.0	93.4
no2	93.0	83.5	95.8
no3	98.3	86.9	97.4
no4	98.0	83.7	98.5
no5	97.5	86.9	96.9

表 12. 日経サイエンスの適合率

9.1 結束チャートの自動生成結果の考察

この結果より以下の考察が得られる。

1. 方法 1 の正解率が 100% にならなかった理由としては、形態素解析の際、正しく語が切り出せない場合があったことと、手作業の際のデータ処理のミスが考えられる。

2. 方法 2 の正解率が低い原因としては、段落内頻度だけでは、分類を決定するための情報が少な過ぎることが考えられる。特に天声人語ではこの傾向が顕著である。

3. テキスト頭からの頻度をとった方がデータ量が多いため結果的に良い結果が得られた。

4. 方法 2, 方法 3 のこの正解率の差は、ほとんどが多義語選択に関する精度によるものである。

5. 方法 3 について 9 割を越える正解率が出たということは、事前調査から得られた結果及び観察が有用なものであったことを示している。

10 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法が考えられる。

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. 文章の要約に応用する。

4. 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、結束チャートの自動生成の際の多義語の処理方法を今回挙げた方法以外にさらに検討するとともに、ここにあげた結束チャートの利用法の検討を進める予定である。

謝辞

「角川類語新辞典」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。

参考文献

[1] 佐々木一朗、増山繁、内藤昭三 : 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, 情報処理学会第 46 回全国大会講演論文集 (3) ,pp.3-177 ～ 178, 1993.

[2] 大野晋、浜西正人 : 角川類語新辞典 角川書店, 1981.

[3] 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.

[4] 伊藤俊一、阿部純一 : 接続詞が持つ制約と連文の結束性, 情処研究報告 NL-86-4, pp.1 ～ 8, 1991.

[5] 佐野、槇、森、中川 : 談話の結束性を考慮した比喩理解過程の解析 (1) - 結束性要因の抽出 - , 情報処理学会第 4 4 回全国大会講演論文集 (3), pp.3-213 ～ 214, 1992.

[6] 槇、佐野、森、中川 : 談話の結束性を考慮した比喩理解過程の解析 (2) - 結束性要因充足としての比喩理解 - , 情報処理学会第 4 4 回全国大会講演論文集 (3), pp.3-215 ～ 216, 1992.

[7] 工藤育男 : 文と文の結束性を捕らえるための知識, 情処研究報告 NL-76-7, pp1 ～ 8, 1990.

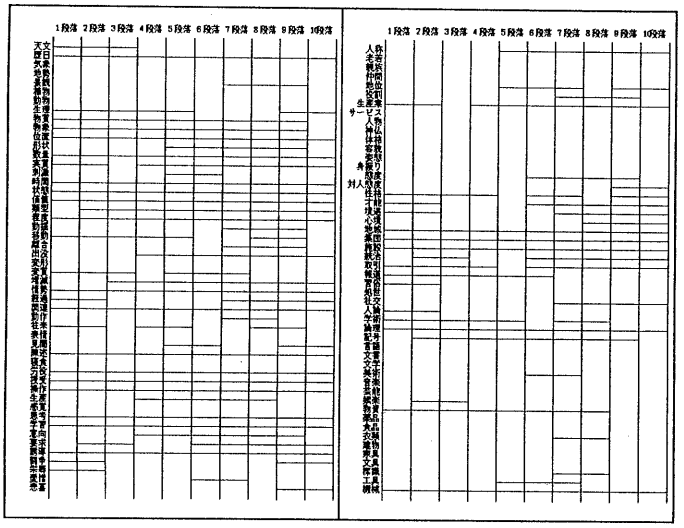

図1 手作業によるチャート (サイエンス)

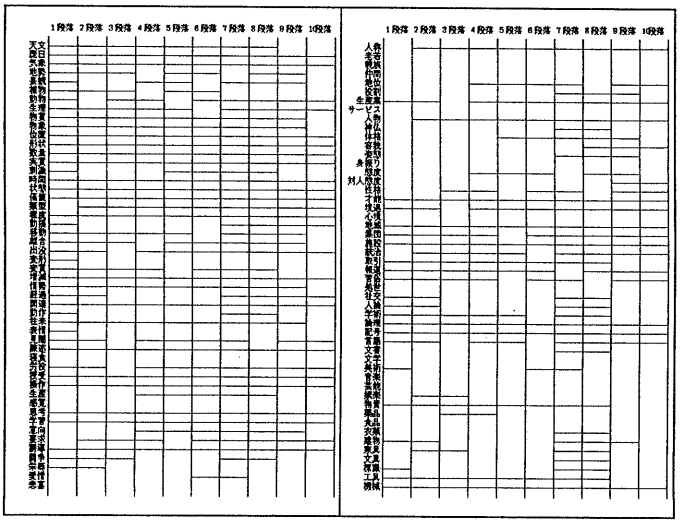

図2 方法1によるチャート (サイエンス)

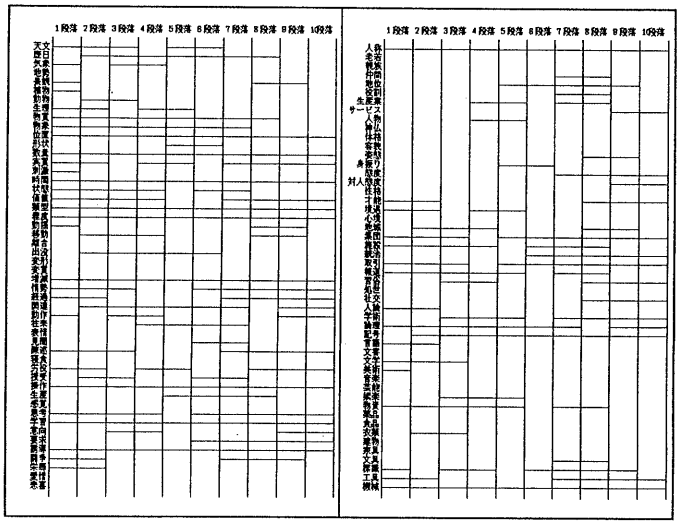

図3 方法2によるチャート (サイエンス)

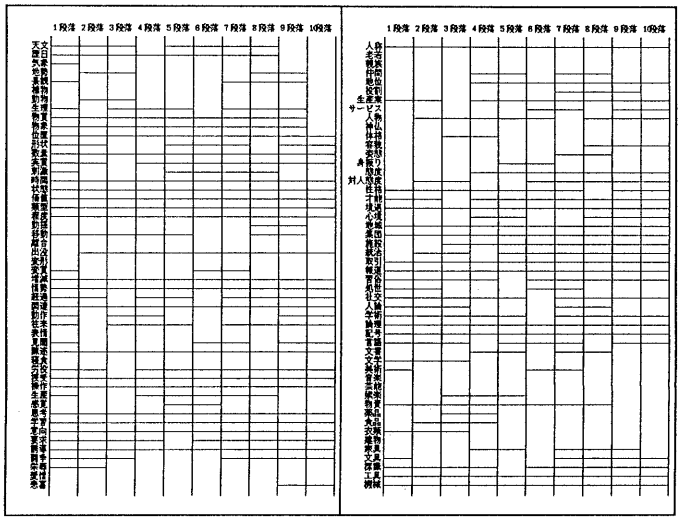

図4 方法3によるチャート (サイエンス)