語の重要度を考慮した談話構造表現の抽出

斎藤尚子　　横山晶一
山形大学大学院理工学研究科　山形大学工学部

1. はじめに

コンピュータの発展に伴い、電子化された文章が多く存在している。その中から必要な情報を抽出するために、キーワード検索などを行う。しかし、その場合でも、不必要な情報が混在することが多い。このような状況で、大きな話のまとまりである談話の構造を計算機で捉えることができれば、効率よく必要な情報だけを抽出できるようになる。

本研究では、文章の内容を計算機で捉えるための一手法として、談話構造表現抽出のためのネットワークを提案する[1]。このネットワークによって、談話中の語句の結びつきと談話構造を同時に捉えることができる。

具体的には、主題・焦点・関連情報を抽出し、これらに対して、文間における結びつきや表題との関係などから重要度を決定する。文間の結びつきは「日本語語彙大系」[2]の属性番号によって決定する。最後に、情報の重要度と結束性をネットワークで表現する。

作成したネットワークは、重要な語句に高い点数を与えるとともに、談話構造をよく反映したものとなっている。評価のために、人間の判断との簡単な比較も行い、よい結果が得られた。それについても述べる。

2. 主題・焦点・関連情報の抽出方法

本研究では、主題・焦点・関連情報という3つの要素を重要情報として抽出する。主題・焦点に関してはさまざまな定義がある[3,4]が、本研究では、以下のように定義する。

主題：前述された既知の情報
焦点：その文で新しく導入された情報
関連情報：主題・焦点の補足や、言い換えに用いられる情報

機械処理をする上で問題となる曖昧性の排除と、主題・焦点の効率的な抽出のために、次の前提条件を設ける。

＜前提条件＞

1）対比の「は」を含むような文は対象外とする。
2）格関係を持たない文は対象外とする。
3）曖昧性のない構文解析木が作られているものとする。
4）時相名詞・副詞は主題・焦点としない。

主題・焦点抽出においては、文を述部の形から、動詞文」・「形容詞文」・「名詞文」の3種類に分け、種類に応じて抽出方法を変える。

本研究で用いる参考資料[5]は、表題といくつかの段落題目がついている。段落題目が、段落で述べられている重要語句を表している場合もあるが、そうでない場合もあるので、本研究では段落題目から得られる情報は用いない。また、表題の直後に位置する文を第1文、それ以外を第2文以降として区別して扱う。

以下に処理の概略を示す。

1）主題を表す「は」が存在
　主題=「は」の前にある語
　動詞文：焦点=「が」格 なければ 文末の必須格 なければ 述語
　形容詞文：焦点=「が」格 なければ 述語
　名詞文：焦点=述語名詞
2）主題を表す「は」が存在しない
　動詞文：主題=第1文…表題中の 「は」の前にある名詞 なければ 最後の体言

第２文以降…前文の主題を補完

焦点=「が」格 なければ 文末の必須格

なければ 述語

形容詞文：主題= 第１文…表題中の 「は」の前にある名詞 なければ 最後の体言

第２文以降…前文の主題を補完

焦点=「が」格 なければ 述語

名詞文：主題= 第１文…表題中の 「は」の前にある名詞 なければ 最後の体言

第２文以降…前文の主題を補完

焦点=「が」格 なければ 文末の必須格

なければ 述語

3)「は」なし、「～を･･･と＋動詞（例：「いう」、「する」）」の場合（動詞文として分類）

動詞文：主題=「研究者」を補完（科学論文ではこのケースが多い）

焦点=「を」格

関連情報は、語句の意味から抽出するのではなく、形式によって抽出する。例えば「～は…と呼ばれる」という形式の場合、「…」を「主題（～）に対する関連情報」として抽出する。関連情報の種類としては、以下のものがある。

①主題に対する関連情報の抽出

1）主題に対する関連情報

2）主題の修飾部中にある関連情報

②焦点に対する関連情報の抽出

1）述部によって導き出せる関連情報

2）焦点の修飾部中にある関連情報

③関連情報と関連する関連情報

3. 語の重要度

主題・焦点・関連情報について、重要度を計算する。主題・焦点・関連情報は、抽出された段階では０点である。重要度計算にあたっては、「文内における主題・焦点・関連情報の重要度」、「文間での３要素の結びつき」、「接続詞」、「表題との関係」という４項目に着目し、条件によってことなる点数を加点する。最後に４項目の点数を主題・焦点・関連情報それぞれにおいて合計して、重要度を決定する。

以下に重要度のつけ方について述べる。

①文内における主題･焦点の点数

主題に＋３、焦点に＋２する。

関連情報は主題･焦点と同様の重要度を持つと考え、関連する主題・焦点と同じ点数をつける。

②文間における主題･焦点の点数

前後の文（第ｎ文と第（ｎ＋１）文）における主題・焦点の結びつきから、表３のように点数を加点する。

表３　文間における主題・焦点の点数

		(n+1)文	
	同じ部分	主題	焦点
n文	主題	(n+1) 文の主題に＋3	点数なし
	焦点	n文の焦点に＋2 (n+1) 文の主題に＋2	n文の焦点に＋2 (n+1) 文の焦点に＋2

		(n+1)文	
	同じ部分	主題の修飾部	焦点の修飾部
n文	主題	n文の主題に＋2 (n+1) 文の主題に＋1	n文の主題に＋2 (n+1) 文の焦点に＋1
	焦点	n文の焦点に＋2 (n+1) 文の主題に＋1	n文の焦点に＋2 (n+1) 文の焦点に＋1

第ｎ文の主題・焦点の修飾部と第（ｎ＋１）文との関係に対しては、加点しない。

③接続詞による点数

「説明」（例：「つまり」、「すなわち」）の接続詞がある場合、文中の主題・焦点・関連情報に＋３する。

「転換」（例：「さて」、「ところで」）の接続詞には点数を加えない。それ以外の接続詞の場合、文中の主題・焦点・関連情報に＋２する。

④表題との関係

主題・焦点・関連情報の語句の中に表題と同じ語句が出現している場合、その語句に＋３する。

⑤重要度算出方法

①～④までの点数を合計したものが、主題・焦点・関連情報の点数となる。ただし、談話構造表現ネットワーク作成の際、以下のような場合においては、点数を変化させる。

1)「関連情報の点数 ＜ 関連している主題または焦点の点数」の場合、関連情報の点数を、関連している主題または焦点の点数と同じになるよう上げる。

2)「関連情報の点数 ＞ 関連している主題または焦点の点数」の場合、関連している主題または焦点の点数を、関連情報の点数と同じになるよう上げる。

3)焦点が、主題の関連情報としても抽出される場合、焦点かつ主題の関連情報である語句の点数は、

・主題よりも点数が低い場合には、主題と同じ点数になるよう上げる。
・主題よりも点数が高い場合には、主題の点数を焦点かつ主題の関連情報と同じ点数になるよう上げる。

4. 談話構造表現ネットワーク

談話構造表現ネットワークでは、主題・焦点・関連情報の文内・文間における結びつきが表現されている。図1に、文内における主題・焦点・関連情報の表現方法、図2に文間における主題・焦点・関連情報の関係の表現方法を示す。

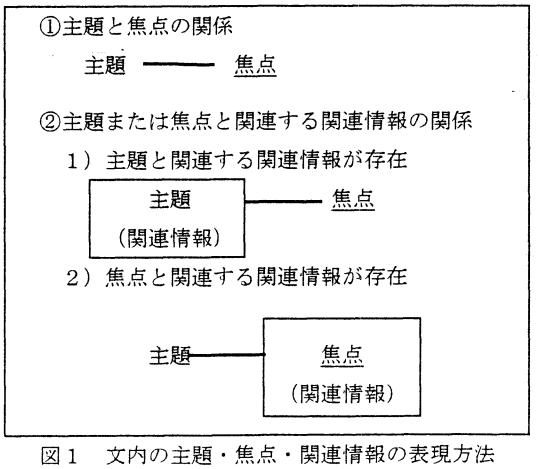

図1　文内の主題・焦点・関連情報の表現方法

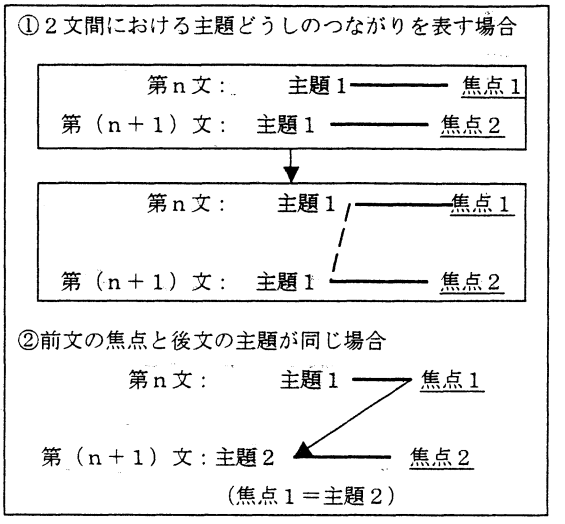

図2　文間の主題・焦点・関連情報の関係の表現方法

5. 談話構造表現ネットワーク作成結果

実験として、Newton1991年2月号に掲載されている「アルツハイマー型老年痴呆」の談話構造表現ネットワークを作成した。図3にその結果の一部を示す。この談話構造ネットワークから、

1）主題と焦点、関連情報として抽出された情報の重要度
2）同じ主題である文が何文続いているかということ
3）文間における主題・焦点・関連情報のつながり

を知ることができる。

重要度が高い情報は、主題として何文にもわたって出現している情報である。例えば、8文目の主題である「老人痴呆」は、14文目まで続いており、14文目の主題「老年痴呆」の重要度は24となる。

この談話構造表現ネットワークにおいて、重要度が高い語句は「老人痴呆」、「異常構造物」、「神経細胞」、「脳」、「人間」、「65歳前（初老期）におこるアルツハイマー病と、65歳以降（老齢期）におこる原因不明の老年痴呆」、「この老年痴呆の方（関連情報：アルツハイマー型老年痴呆）」が挙げられる。

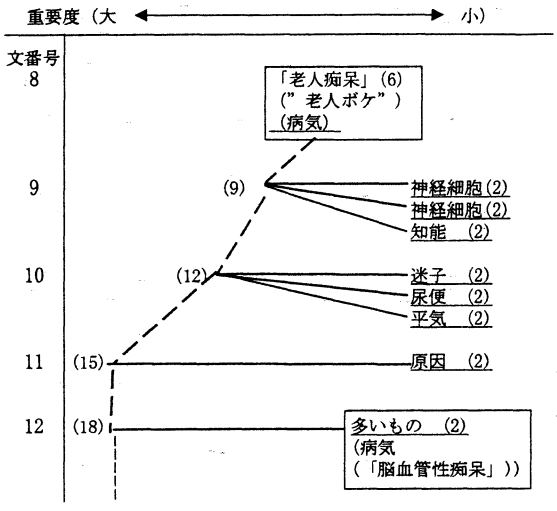

図３「アルツハイマー型老年痴呆」の談話構造ネットワーク（一部）
（主題は網掛け、焦点は下線を引いて表す。）

この談話構造解析ネットワークを評価するために、８人の被験者に本文を読んで、その中から重要な語句を３つ抽出してもらった。その結果と照らし合わせると、４人以上の被験者によって抽出された語句の中に「老年痴呆」、「アルツハイマー型老年痴呆」、「異常構造物」、「神経細胞」が含まれている。よって、本研究において重要情報として抽出された語句は、人間にも重要だと認識されていることになる。

この談話表現ネットワークでは、主題の連続性を知ることができるので、主題の連続性を用いて段落をつくり、その段落中から重要文を抽出して談話の要約を作るということに活用できる。また、談話の流れを捉えることで、談話の内容が何を中心に述べられているのかを知ることができる。それを用いて必要な情報のみを厳選するための手段として用いることが可能になる。

6. おわりに

本研究では、談話構造を理解するための第一段階として、談話構造表現ネットワークの構築を試みた。

今回用いた資料では良い結果が得られたが、このネットワークが他の種類の談話に関しても通用するかは、今後検討の予定である。

また、点数のつけ方について、日本語の談話の形式として「起・承・転・結」がある。人間による重要度の抽出においても、談話の最後の部分にある語句を重要だとする人が多かった。そこで、今後、「起・承・転・結」を考慮した点数付けについて検討していく必要がある。

参考文献

[1] 斎藤尚子：語の重要度を考慮した談話構造表現に関する研究、山形大学大学院理工学研究科修士学位論文（2000）

[2] 池原悟、宮崎正弘、白井諭、横尾昭男、中岩浩巳、小倉健太郎、大山芳史、林良彦：日本語語彙大系 CD-ROM 版、岩波書店（1999）

[3] 清水一澄、横尾英俊：日本語理解システムのための視点抽出と照応解決、情報処理学会論文誌、Vol.36、No.2、pp.236-246（1995）

[4] 野田尚史：新日本語文法選書１「は」と「が」、くろしお出版（1996）

[5] ニュートンプレス：SCIENCE BOX、Newton、1991、1992 年版