談話構造タグ付け半自動化についての一考察

竹内和広、松本裕治
奈良先端科学技術大学院大学 情報科学研究科
e-mail:{kazuh-ta, matsu}@is.aist-nara.ac.jp

1 はじめに

近年、計算機による文書要約等のテキスト処理において談話構造を利用する技術への期待が高まっている。このように談話構造を利用した計算機処理を実現するために、その基礎データとして、書き言葉における談話構造解析済みテキストを効率的に蓄積することが課題となってきている。本研究では日本語の新聞報道記事に対して修辞構造理論 (RST) のサブセットを用いて複数の被験者による構造解析タグ付け実験を実際に行い、テキストの構造的な要因とタグ付けの信頼性との関係を考察した。また、そこで得られた知見をもとに、 現在、比較的高い解析精度を達成している自然言語解析モジュールを複合的に用いて、談話構造タグ付けがどの程度まで半自動化が可能か検討を行った。

2 談話構造タグ付け実験

談話構造をタグ付けするために RST(rhetorical structure theory [Mann & Thompson 1987] を参考に表1のような修辞タグ付け体系を再定義し、被験者3人に対する心理実験をおこなった。表中の修辞関係名に英大文字で記したものが Mann らのオリジナルの修辞関係名である。

タグ付けの対象は日本経済新聞記事の報道記事を用いた。その際に、政治、経済、文化等の分野は問わずに、95年1月から6月までと12月の記事から記事の長さが10文から30文程度で構成されるものを無作為に32記事 (合計500文) 選択した。報道記事を選んだ理由は他の記事に比べ、目的が限定され、用いられる修辞関係が少ないことから被験者のゆれが少ないと考えたからである。

タグ付け作業は図1のようなタグ付けエディタを用いて行った。タグ付けエディタにおいて、文と文の関係付けはマウスで関係元の文と関係先の文を順にクリックする操作で行うことができ、関係付けられた2文間には関係付け矢印が引かれる。関係付け矢印をクリックするとメニューが画面上に表示され、そこから関係種類を選択することができる。選択された関係種類は関係付け矢印の色分け表示に反映される。将来的には、このようなタグ付けエディタにタグ付け支援機能を埋め込む予定である。

タグ付け体系の教示は簡単なマニュアルを作成して行った。それを被験者に読んでもらった後に、練習問題を解いてもらった。練習問題を解く際には質疑応答を口頭で行った。その際、関係付けの制約として、テキスト中の一文を除くすべての文に対して、それぞれに一番関係が深いと思われる文をひとつだけに関係付けることを指示した。また、関係先の文が関係元の文に循環して結び付く構造は禁止した。なお、形式段落は画面上に表示するが、タグ付けに関わる特別な指示は行

表 1: 整理した修辞関係

修辞関係	修辞関係の説明
bac(BACKGROUND)	背景状況を提供する
ela(ELABORATION)	補足・詳細説明
res(RESATEMENT)	言換え・要約をする
seq(SEQUENCE)	連続事態の提示
con(CONTRAST)	対比、対立
quo	引用
「???」	不明 (予備用)
nil	根 (1記事に1文のみ許す)

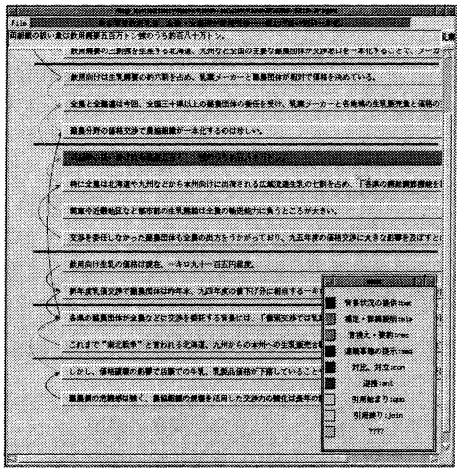

図 1: 修辞構造タグ付けエディタ

わなかった。

3 タグ付け実験の結果

このような実験に対し、関係先、および関係先の一致率は表2のようになった。評価には [Carletta et al.1997] らが導入を提案した Kappa 統計値という指標を用いた。Kappa 統計値の算出は観測一致率を P(A)、偶然一致率を P(E) とする以下の式を用いた。

$$Kappa値 = \frac{P(A) - P(E)}{1 - P(E)}$$

Kappa 値による信頼性の評価において、1) 0.41 から 0.60 では適度 (moderate) の一致、2) 0.61 から 0.80 では内容のある (substantial) 一致、3) 0.81 から 1.00 では完全に近い (near perfect) 一致、という基準があるとされる。本実験の結果は、この基準で判断すると、被験者間の関係先、関係種類の一致率ともにある程度の傾向を示すものの、それほど高い値ではない。

そこで、タグ付け一致の傾向をさらに詳細にみるために、ある文が自分を中心として何文前の文に関係するか（後方の文の場合は負数）を関係距離として定義し、各被験者の関係付けに対する比較一致率を調べたところ、表3、表4、表5のよう

表 2: 関係先・関係種類の一致率

一致の種類	観測一致率	偶然一致率	Kappa 値
関係先	0.63	0.11	0.58
関係種類	0.69	0.46	0.43

表 3: 被験者 A に対する比較一致率

距離	関係数	被験者 B	一致率	被験者 C	一致率
根	32	28	87.5%	29	90.6%
1	309	210	68.0%	238	77.0%
2	69	31	44.9%	15	21.7%
3	33	12	36.4%	13	39.4%
4	15	4	26.7%	4	26.7%
5	7	2	28.6%	3	42.9%
6 以上	23	5	21.7%	3	13.0%
-1	10	7	70.0%	2	20.0%
-2	1	0	0.0%	0	0.0%
-3 以下	1	0	0.0%	0	0.0%

表 4: 被験者 B に対する比較一致率

距離	関係数	被験者 A	一致率	被験者 C	一致率
根	32	28	87.5%	31	96.9%
1	255	210	82.4%	233	91.4%
2	68	31	45.6%	24	35.3%
3	20	12	60.0%	14	70.0%
4	22	4	18.2%	12	54.5%
5	13	2	15.4%	7	53.8%
6 以上	55	5	9.1%	17	30.9%
-1	33	7	21.2%	11	33.3%
-2	2	0	0.0%	0	0.0%
-3 以下	0	0	—	0	—

表 5: 被験者 C に対する比較一致率

距離	関係数	被験者 A	一致率	被験者 B	一致率
根	32	29	90.6%	31	96.9%
1	301	238	79.1%	233	77.4%
2	41	15	36.6%	24	58.5%
3	35	13	37.1%	14	40.0%
4	17	4	23.5%	12	70.6%
5	16	3	18.8%	7	43.8%
6 以上	41	3	7.3%	17	41.5%
-1	16	2	12.5%	11	68.8%
-2	0	0	—%	0	—%
-3 以下	1	0	0.0%	0	0.0%

な傾向がみられた。表中の比較一致率は以下のような式で算出し、単位を(%)で示した。

$$\text{比較一致率} = \frac{\text{2人の被験者の当該関係距離で一致する関係数}}{\text{比較される被験者の当該関係距離における関係数}}$$

この結果から、どの被験者も関係距離1の関係が非常に多く、関係距離が遠くなってゆくに従って、その数が減って行くことがわかる。また、ほかの被験者との比較一致率も関係距離1の関係がもっとも高い。このような関係距離における一致率の傾向から、実験ではすべての組合わせの文関係を許したにもかかわらず、隣接文関係についての人間の作業者の一致率が高いことがわかった。そこで、これらの隣接文関係を半自動的にタグ付けできないかを次節で検討する。

4 決定木学習および学習で用いる属性

前節のような結果をうけて、隣接文関係が存在するか否かを自動的に分析する規則を決定木学習する試みを行った。

決定木学習は[Quinlan 1992]によるモジュールであるC4.5を用いた。決定木学習の学習データとして属性は、隣接する2文間の特徴付けをすると考えられる以下のような言語情報を用いて整理した。

これらの属性は、現在ある程度の精度を達成している自然言語処理モジュールを積極的に利用して元テキストから抽出する。その際の解析誤りは人手によって修正した。

- 文末表現属性
  - 文末表現を書き手の意志の介在する表現によって分類
    モダリティ表現のタイプを分類
  - 文法的な表現を分類
    文末が名詞で終わる、「だ」形で終わる、「ている」形で終わる等を分類。
- 主語、主題属性
  あらかじめ、かかりうけ解析処理[藤尾 & 松本1997]を行い、文末にかかる「が」格「は」格「も」格の存在を利用。同時にそれらの表層格すべてが存在しないという情報も利用し、これらの表現の存在の有無を値とする。
- 文頭表現の属性
  日本語形態素解析システム『茶筌』の辞書から接続表現を抽出し、15種類に分類した。
- 同一指示対象を持つ名詞句属性
  名詞句の指示対象同定モジュール[山根 1999]を用いて、当該の2文間にどれだけ共通の指示対象をもつ名詞句が存在するかを属性として用いた。
- 形式段落属性
  当該の2文が同一形式段落に入るか否かを属性として用いた。

5 決定木学習によって得られた規則

前節で整理した属性を用いて決定木学習した結果を表6および表7に示す。表に示した正解率の算出は、学習で得られた規則によってテストデータを解析した結果とそれぞれのタグ付けされた値とが一致した時を正解とし、以下のような式で計算を行った。

$$\text{正解率} = \frac{\text{正解した数}}{\text{機械学習した規則によって解析した数}} \times 100(\%)$$

表6は、学習データとしてそれぞれの被験者のタグ付けデータを用いて学習した結果である。テストデータとしては、学習データと学習した被験者以外のタグ付けデータをそれぞれ用いた。なお、学習後の決定木の枝刈りは行っていない。この結果から、それぞれの被験者のタグ付けデータから得られた規則がある程度のゆれはあるものの他の被験者がタグ付けしたデータに対しても一般性を持つことが示されている。

また、表中には前節で整理したすべての属性を用いて学習した例とそれらから形式段落属性を省いたものから学習した例の2通りの結果を示してある。これは予備的な実験において、形式段落属性がもっとも学習データをよく分割していたため、参考用に提示する。

表7は、被験者がタグ付けしたデータすべてをまとめて検証した結果である。評価はすべてのデータを5つに分け、そのうち4つを学習に用いる交

表 6: 各被験者のタグ付けデータから学習

学習データ	テストデータ	用いた属性 / 正解率	
		すべての属性	段落属性なし
被験者 A	学習データ	84.0 %	82.2%
	被験者 B	70.6 %	64.2%
	被験者 C	68.8 %	65.6%
被験者 B	学習データ	86.0 %	79.4%
	被験者 A	69.6 %	67.0%
	被験者 C	80.6 %	74.8%
被験者 C	学習データ	88.0 %	81.8%
	被験者 A	69.4 %	69.6%
	被験者 B	76.2 %	71.2%

表 7: 全員のデータすべてから学習

枝刈り	用いた属性 / 正解率	
	すべての属性	段落属性なし
枝刈りなし	67.2 %	60.8 %
枝刈りあり	74.7 %	63.7 %

差検定を行い、正解率はその平均値を示した。表中で最も正解率の高い結果である 74.7% はすべての被験者に対して一般的に適用できる正解率であると考えられる。

このような決定木学習で得られた正解率を評価するために、テキスト中の文すべてが前文に関係するとする単純な規則と比較すると、すべて前文に関係するとした規則の正解率は、被験者 A,B,C に対してそれぞれ、66.0%, 54.5%, 64.3% にとどまるため、本研究で得られた規則が隣接関係の自動解析をある程度可能にしていると言える。

今回行ったどの実験においても、学習に用いた属性の中で形式段落の属性が最もよく学習例を分割した。しかし、形式段落の設定は書き手によって基準にゆれがあり、このような形式段落の属性をどのように扱って行くかは今後の課題である。

6 おわりに

本研究では、現段階である程度の解析精度を達成している自然言語解析モジュールを複合的に用いて、人間の作業者間である程度一致が見られる関係については、計算機でも半自動的な解析が可能なことを示した。

本研究で試みた手法は、談話構造タグ付けのためのモデルとは別に言語情報の蓄積を行うことができることが利点で、談話構造に関係すると思われる言語情報を精度よく蓄積することによって、新しい談話構造モデルにも柔軟に対応することができ、談話研究における基礎データを効率的に蓄積することができるのではないだろうか。

謝辞

新聞報道記事を本稿の実験の対象として使用させていただきました日本経済新聞社に感謝します。また、本研究をはじめるきっかけを与えて下さいました (株) 日立製作所 基礎研究所の 野本忠司氏とタグ付け実験に協力してくださった方々にこの場を借りて御礼申し上げます。

参考文献

[藤尾 & 松本 1997] 藤尾正和、松本裕治. "EDR 括弧付きコーパスを利用した、統計的日本語係り受け解析" EDR 電子化辞書利用シンポジウム 論文集, pp.49-55(1997).

[松本ほか 1997] 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明 . "日本語形態素解析システム『茶筌』 version 1.0 使用説明書" NAIST Technical Report, NAIST-IS-TR97007(1997).

[山根 1999] 山根洋平. "文章中の名詞間照応関係の同定" 奈良先端科学技術大学院大学情報科学研究科修士論文 (1999).

[Carletta et al. 1997] Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics, Vol.23, No.1*, pp.13-31(1997).

[Mann & Thompson 1987] William C. Mann, Sandra A. Thompson. *Rhetorical Structure Theory: A Theory of Text Organization.* Technical Report ISI/RS-87-190, ISI Reprint Series(1997).

[Quinlan 1992] J. Ross Quinlan. *C4.5: Programs for Machine Learning.* The Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann (1992).