自然言語処理 133－9
（1999．9．10）

自動要約を視野にいれたテキスト構造解析実験

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

近年、計算機による機械要約等のテキスト処理においてテキスト構造を利用する技術への期待が高まっている。このようにテキスト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、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。本研究では修辞構造理論を簡略化したテキスト構造解析タグ付け体系を試作し、日本語の新聞報道記事に対し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、被験者間の一致の傾向を観察した。また、被験者において一致の見られる部分については、決定木学習を用いて言語情報からどの程度その部分を推定できるかを検討した。

An Empirical Analysis of Text Structure as a Basis for Automated Text Summarization

Kazuhiro Takeuchi and Yuji Matsumoto
Graduate School of Information Science, Nara Institute of Science and Technology
e-mail:{kazuh-ta, matsu}@is.aist-nara.ac.jp

Methods for automated text summarization have gained increasing importance with the rapid growth of machine-readable texts. To develop such a method, it is useful to know the structure of a text, since the structure shows the relative importance of the consitiutent sentences in the text.

In this paper, we first report some psychological experiments to analyze the structure of texts using a simplified version of RST (Rhetorical Structure Theory). The result reveals that human subjects put extra importance on the relationship between two adjacent sentences in the texts.

We then apply decision tree learning to automatically learn those adjacent pairs in which human subjects find a direct relationship. As a result, we obtained approximately 70% accuracy by the learned decision tree.

1 はじめに

近年の電子化テキストの増大により、計算機を利用して効率的にテキストを処理する技術に対して期待が高まっている。このような中で、Ono et al[6]、Marcu[8]、比留間ら[2]の研究のようにMann & Thompsonが提案したRST(Rhetorical Structure Theory)[7]を用いてテキストの構造解析を行い、その情報をもとに自動要約を行う試みがなされている。RSTはテキストを部分領域に分け、ある部分領域が他の部分領域に対して対照関係や詳細説明関係にあるといった部分領域間の関係的意味を用いてテキストの構造を階層的に分析し、テキストの首尾一貫性を説明しようとする枠組である。

このようなテキストの構造情報を要約に用いる主な目的を整理すると、単にテキスト上の重要部分を抽出するだけではなく、重要な箇所に関連する箇所も抽出し、また、どのような文章の構成の中でその箇所が重要であると判断されているかを説明づけるといったことが挙げられる。

しかし、RSTの理論はこのような特定の応用

を目指して提案されたものではなく、テキストに対する一般的な解析を目指したものであるため、RST に基づいたテキストの解析には解析者によってあいまい性があることが指摘されている。また、自動要約を視野にいれてテキスト構造解析を考えた場合、自動的にテキストの構造を解析するにはRST で定義されている修辞関係の種類は多く、修辞関係を自動要約に有効に利用する方法についても課題が残っている [9]。

そこで、本研究では、テキストの部分領域間の関係性の強さと、重要性の相違という観点を用いて、人間がどのように日本語のテキスト構造を分析するかを調査した。また、人間によって分析が一致する部分については、いくつかの言語情報をもとに決定木学習を行い、人間の分析をどの程度推定できるか検討した。

2 テキスト構造タグ付け体系の定義

2.1 RST の単純化

本研究では、RST を参考にタグ付け体系を試作し、そのタグ付け体系を用いて人間にテキストの構造を解析してもらう方針をとる。

RST では修辞関係を客観的に定義しようとしているものの、Moore & Pollack[10] や Moser & Moore[11] が指摘するように、部分領域間に選択された修辞関係の種類によっては同時に多重解析が可能であり、テキスト中の修辞関係により結ばれた 2 つの部分領域のうちその修辞関係についてどちらをより内容的に中心的とみなすかにより定義される RST の関係方向 (本研究ではより中心的な側を関係先，他方を関係元と呼ぶ) が逆になってしまう例があることが指摘されている。

本研究では、テキストの構造を決定する上で、修辞関係よりも関係方向の決定が重要と考え、まず、最も関係が深い部分領域対の同定とその部分領域間の関係方向の選定を行い、その後に関係種類の選択を行う解析手順を考えた。また、RST で定義された修辞関係すべてを用いるのではなく、まずはおおまかな関係的意味の区別を行っておき、将来的にその関係を詳細化する方向性を考えている。

本研究における関係方向の定義は、部分領域間の重要度の相違をもとにする。自動要約の研究では、例えばある特定のテキストに集中して出現する語句はそのテキストにおいて重要な語句であるといったような統計的な情報を利用して、重要文や重要語句を選定し、抄録として抽出する手法が提案されている [1]。つまり、要約において、テキストの部分領域同士を比較した場合に、どちらの部分領域がより重要であるかという判断は自動要約を行う上で本質的に必要な情報であると考える事ができる。そこで本研究では関係方向の選定として、関係を持つ 2 文のうちどちらがより重要と考えるかを被験者に判断させ、一致の傾向を観察しようと考えた。

部分領域の最小単位については、文を想定する。RST の元々の定義ではテキストの部分領域は日本語の複文における節に相当するものを仮定している。しかし、複文における従属節は、文内で関係付けが解決されることが期待できるため、テキスト構造解析の枠組においてそれらを二重に解析する必要はないと考え、文と文同士の関係の関係に焦点を合わせることにした。このように、本研究でも、他の多くの先行研究が想定するように、部分領域の基本単位である文の関係付けから、テキストが階層的に分析されることを想定する。また、本実験では文間の関係付けにおいて循環の構造は禁止するものの、テキスト上の連続的な文のならびに対して関係付けが交差するような構造は禁止しない。つまり、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許す。

2.2 テキスト構造解析タグ付け体系

前節で検討した内容を元に、テキストの構造解析用のタグ付け体系を試作する。

被験者が実際にテキスト中のそれぞれの文に対して行う選択を以下の手順に制約する。

1. 当該文と最も関連の深い文の選択
2. 1 で選んだ文とどちらが重要であるかの選択
3. 関係種類の決定

関係の種類については RST の修辞関係を参考にいくつかの関係種類を採用した。その際、重要性の観点において明らかな差異があるものと、重要性の序列をつけにくい並列的な関係との区別を最も基本的な区別とし、計 6 種類の関係種類に整理した。今回の実験では、テキストがどのような構造に解析されるかを第一の目的とするため、導入したおおまかな関係種類に対する詳細化については別の機会に検討する。

3 テキスト構造解析タグ付け実験

3.1 実験の内容

試作したテキスト構造解析タグ付け体系を用いて、実際に被験者3人にテキストを解析させる心理実験をおこなった。

タグ付けの対象は日本経済新聞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、記述の目的や用いられる修辞関係が限定され、被験者による解析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作成して行った。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。このような選択の過程は、タグ付け体系の設計の際に仮定した選択の手順にしたがっており、このようなタグ付けエディタを用いる事により、被験者の作業を統制する。また、将来的には、このようなタグ付けエディタに自動的なタグ付け支援機能を埋め込み、データを効率的に蓄積することも考えている。

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。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っていない

3.2 タグ付け実験の結果

実験の評価は複数の被験者で解析が収束するか否か、すなわち被験者がタグ付けをした内容の一致をもとに評価する。

まず、総合的な評価を行う。タグ付け体系の信頼性の評価についてはCarlettaらの談話研究[5]において用いられたKappa統計値という指標を利用した。Kappa統計値の算出は観測一致率をP(A)、偶然一致率をP(E)とする以下の式で定義される。

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

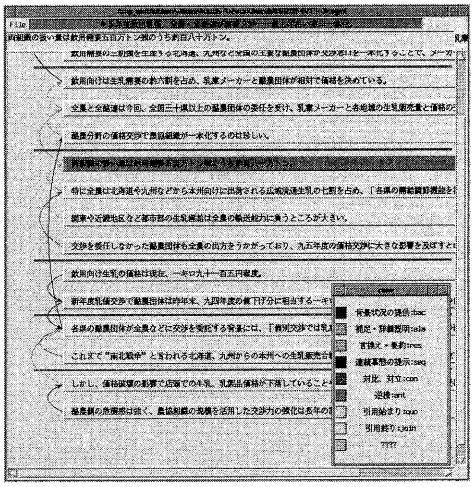

図1: 修辞構造タグ付けエディタ

表1: 関係先の一致率

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

この指標を用いて、関係先の一致に対する評価を行うと結果は表1のようになった。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ではかなりの(substantial)一致、3) 0.81から1.00では完全に近い(near perfect)一致という基準が提案されている。この基準で判断すると、被験者間の関係先一致にはある程度の傾向が存在するものの、必ずしも高い値ではないことがわかる。

そこで、タグ付け一致の傾向をさらに検討するために、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2、表3、表4のような傾向がみられた。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。

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

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

表 2: 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
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 3: 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
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	—

係を許したにもかかわらず、より遠く文との関係と比較して、関係距離 1 の関係についてはよく一致することがわかった。

今回タグ付けされた全 500 文の中で関係先を多数決 (被験者は 3 人のためタグ付けの値が 2 者以上で一致するもの) によって決定できない例は 45 文であった。つまり、1 記事あたりの平均文数 15 文のうち平均 1.4 文以外については多数決で関係先を決定でき、構造が解析できることがわかる。また、それぞれの記事において構造木の根の一致率は高く、タグ付けを行った 32 記事のうち 3 人の被験者すべてが別々の文を根とした事例はなく、28 記事については三者一致で第一文を根としている。さらに、3 者一致で関係先が一致する文は根を除くと 222 文存在したが、そのうち 194 文が関係距離 1 の関係だった。

表 4: 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
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%

4 言語情報との関係

タグ付け実験の結果分かったこととして、隣接文間が関係するか否かと言う関係距離が 1 で定義される関係、すなわち隣接関係については人間によって比較的一致する事がわかった。

直観的な観察によれば、このような人間によって一致する関係距離が 1 の関係は形式段落の中に多く分布し、形式段落の境界では、関係距離 1 以外の距離で他の文と結び付くことが多い。実際、人間が関係距離を 1 以外とした 539 事例のうち形式段落の境界にあたるのは 380 事例存在し、関係先が関係距離 1 以外の文の 70.5% が形式段落の先頭文である。また、形式段落の先頭文は 474 事例存在するので、形式段落の先頭文の 80.1% が関係先を関係距離 1 以外としている。

形式段落は通常、書き手によるゆれがあるものの、話題のまとまりとして設定される。このことを考慮にいれると、関係距離 1 の関係の分布の傾向は、関係距離 1 の関係は直前の文と記述内容に密接な関係があり、関係距離 1 以外の関係が定義されている隣接文間では局所的な話題境界点となっているのではないかと思われる。

以上のような直観的な観察を統計的に調べるために、テキスト構造解析の中で被験者間での一致が比較的高かった関係距離 1 の関係と言語情報とがどのような関係を持っているかを機械学習の一手法である決定木学習モジュールの C4.5[13] を用いて調べる。具体的には、3 人の被験者が作成した隣接文関係のデータを用いて決定木学習を行い、どの程度の正解率を達成する決定木を獲得できるかを調べることにより、言語情報と人間のタグ付けの関係を考察する。

決定木学習の学習事例およびテスト事例として使用するデータは、タグ付けされた 32 記事の先頭文を除く、468 隣接文間に 3 人がクラス分けした計 1404 件の事例を用いる。分類するクラスはテキスト構造解析実験においてある隣接文間でテキストの終り方向にある文を当該文と呼び、当該文が関係距離 1 で直前の文に関係づけられているとき、その隣接文間を「隣接文関係あり」に分類されているとし、当該文が関係距離 1 以外で他の文と関係づけられているとき (すなわち直前文に関係づけられていないとき) その隣接文間を「隣接文関係なし」として分類する。これら隣接文間の関係

表 5: 属性一覧

属性のタイプ	属性名	属性の値
形式段落属性	形式段落属性	当該文が形式段落の先頭文か否かで分類
統語属性	文頭手がかり語	当該文の文頭表現を表 6のように分類
	直前文の文末タイプ	直前の文の文末表現を表 7のように分類
	当該文の文末タイプ	当該文の文末表現を表 7のように分類
	主語の有無	「は」型、「も」型、「が」型、その他の 4 値に分類
	文頭提題の形式	「は」型、「も」型、その他 の 3 値に分類
記述内容属性	新規 NE の提示	新規の NE が文頭に現れる、それ以外に現れる、現れないに分類
	参照表現の出現位置	NE 参照表現が文頭に現れる、それ以外に現れる、現れないに分類
	話題の字面類似度	字面の類似度を実数値で値とする。(文頭提題がない場合は 1.0)

の有無を分けた事例から、統計的にどのような決定木が生成されるかを観察する。

隣接文関係の特徴を記述するために言語情報を複数の属性として表 5のように整理した。属性は大きくわけて、形式段落属性、統語属性、記述内容属性の 3 つの属性のタイプにわけた。形式段落属性は当該文が形式段落の先頭文か否かを値とする 2 値属性である。以降、統語属性、記述内容属性の順にそれぞれのタイプの属性の概要と属性がとる値について説明する。

4.1 統語属性

言語情報を属性として扱う上での前処理として係り受け解析を行う。係り受け解析は、文中の文節依存関係を解析するもので、統語解析処理の前処理と位置付けられる。本研究では、藤尾らの作成した係り受け解析ソフトウエア [4] を利用して、係り受け解析を行い、解析誤りのある部分については人手によって修正した。

文頭手がかり語属性は上のような係り受け解析の結果を利用し、当該文の文頭に現われる文節から表 6に整理した表現が現れる場合に、それぞれを手がかり表現として属性の値に分類した。

同様に、文末の表現から文のタイプを分類した。先行研究により、単なる事実を叙述する文と書き手の意見や判断が叙述される文等に文タイプを分けることが解析の上で有用であることが知られているが、本実験で対象とした新聞報道文は書き手の意見が表明される文が少ないため、隣接文間における当該文とその直前の文の文末をそれぞれ表 7に分類し、隣接 2 文の文タイプの組で隣接文間の特徴を表現した。

表 6: 文頭手がかり語属性の値

値	分類される文頭表現の例
ITEM	「第一に」、「第二に」等
CONT	「しかし」、「だが」等
RSLT	「結局」、「最後には」等
ELA	「例えば」、「具体的には」等
JOIN	「さらに」、「また」等
REF	「こそあど」で照応
TIME	特定の日時+「、」
CHAR	特殊記号等
NONE	上記以外

主語の有無属性も、係り受け解析で得られた文節間の依存関係利用して値を決定する。具体的には、当該文の文末に係る文節のうち「は」「も」「が」格表示があるものに注目し、文末に係る「は」「も」「が」格表示がある文節をもつ文と、そのような文節を持たない文に分類した。

また、本研究では当該文で文の主題がどのように提示されているかを、文頭提題の形式属性として利用する。図 2のように、文頭もしくは表 6で示したような手がかり表現の次に、「は」格「も」格を含む文節がある文 (その文節の前にその文節に係る連体修飾節がある場合はその連体修飾節を含めて) を文頭提題部分を持つ文と考え、テキストの中で新しい話題を提供する可能性のある文とみなす。属性としては、「は」格で表示された文頭提題部分を持つ文、「も」格で表示された文頭提題部分を持つ文、文頭提題部分を持たないと文という 3 通りの値に分類した。

表 7: 文末タイプ属性の値

値	分類される文末表現の例
MAYBE	「だろう」「に違いない」等
QUO	引用形
TEIRU	「ている」形
DA	「だ」形
TA	上以外のもので過去形
NOW	上以外のもので現在形
000	上以外のもので時制なし

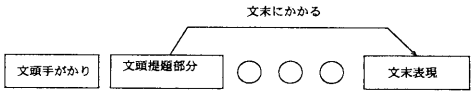

図 2: 文頭提題部分の定義

4.2 記述内容属性

前節で紹介した属性は、主に表層構造から得られる統語情報を、隣接文の特徴を表現する言語情報の属性として整理したものである。

このような属性に対し、隣接する2文でそれぞれどのような内容が記述されているかに関係する属性として、本研究では、テキスト中の記述内容と関連を持つと思われる固有名詞 (Named Entities 以下 NE) の出現と NE を参照する表現、及び字面の類似度を利用した。

NE と NE に対する参照表現は人手により解析した。解析は、まず、ある記事の中から固有名詞を抽出する。この際に、係り受け解析で得られた品詞情報を参考にする。また、本研究で今回 NE として採用したのは、人名、地名、組織名と考えられる固有名詞である。次に、その NE の集合の要素を参照している表現を人手により同定した。

このようにして人手により解析した NE 情報を、隣接文を特徴付ける属性として2つに整理した。一つは新規の NE 表現が現われるか否かの属性で、直前の文の要素に対して参照表現をもたない表現が当該文の文頭提題部に現われる場合、文頭提題部以下に現われる場合、現われない場合の3つの値に分ける。

また、参照表現の出現位置属性として、当該文に直前の文の NE 要素を参照している表現が、文頭提題部に現れる、それ以外の位置に現れる、現われないに分類する。

字面の類似度も、NE 情報と同様に当該文の文頭提題部分で述べられている内容と前文で述べられている内容を比較するために用いる。具体的には、当該文の文頭提題部分と前文をそれぞれ表現するベクトル $\vec{d}$ および $\vec{w}$ をつくり、その2ベクトル間の類似度 sim を以下の式から求める。それぞれのベクトルの要素は、当該文の頭提題部分と全文で使われた文字から、句読点と平仮名を除いた文字がそれぞれの部分で何回使われているかを示すものである。話題の字面類似度属性は、この sim の値を実数値で値とするが、当該文に文頭提題部がない場合は、直前文の話題が継続するものと考え、値を 1.0 とした。

$$\vec{d} = (d_1, d_2, \ldots, d_t), \vec{w} = (w_1, w_2, \ldots, w_t)$$

$$sim(\vec{d}, \vec{w}) = \frac{\sum_t d_t w_t}{\sqrt{\sum_t d_t^2 \sum_t w_t^2}}$$

4.3 決定木学習の結果

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。

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

評価にはデータを5つに分割し、4集合を学習データとして用い、残りの1集合をテストデータとして用いる交差検定を行い、正解率の平均をとった。

また、隣接文関係が連続しない点を話題の境界と考えた評価として、被験者が隣接関係をなしとした事例に対する適合率と再現率を以下の式により算出した。

$$\text{適合率} = \frac{\text{被験者が定義した話題境界と一致した数}}{\text{決定木が話題境界に分類した数}} \times 100(\%)$$

$$\text{再現率} = \frac{\text{被験者が定義した話題境界と一致した数}}{\text{被験者が話題境界を定義した数}} \times 100(\%)$$

表5に示した属性すべてを用いて学習した結果、および形式段落属性のみを用いずに学習した結果を表8に示す。結果では、形式段落属性用いない場合、正解率、適合率、再現率すべてが形式段落属性を用いる場合に比べて低下するため、やはり、形式段落属性が被験者の隣接文関係の有無の選択に強く関係していることがわかる。

このような決定木学習で得られた正解率の比較の基準として、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

表 8: 形式段落属性の用いる場合と用いない場合の正解率 (%)

用いた属性	すべて	形式段落属性なし
正解率	76.0	73.9
話題境界：適合率	71.3	68.3
話題境界：再現率	63.1	59.9

表 9: 形式段落属性を用いない場合の正解率 (%)

用いた属性	形式段落属性以外すべて	SIM 属性なし
正解率	73.9	72.2
話題境界：適合率	68.3	66.2
話題境界：再現率	59.9	56.2
用いた属性	NE 属性なし	SIM 属性及び NE 属性なし
正解率	73.6	70.7
話題境界：適合率	68.3	65.5
話題境界：再現率	58.6	50.3

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% で平均 61.6% である。また、偶然の一致率は 52.3% である。

形式段落属性を用いずにどの程度話題境界が分類できるかについても実験を行った。この実験でも、統語属性タイプの属性はすべて用いている。実験は、話題境界の大きな手がかりとなると考えられる記述内容属性タイプのうち話題の字面類似度属性を SIM 属性、新規 NE の提示属性と参照表現の出現位置属性の 2 つを NE 属性と表記し、それぞれ属性を用いて学習した場合／用いない場合の計 4 通りの学習を行った。結果は表 9に示す。この結果から、記述内容属性が話題境界の判定に有益に働いていることがわかる。また、SIM 属性と NE 属性とでは、SIM 属性である字面の類似度を用いた方が良い正解率を達成しているが、その原因としては、本実験で採用した NE が限定されたものであったことが挙げられる。

形式段落属性を用いない決定木による実際の分類誤りとしては例 1 のようなものがあった。例 1 は形式段落境界である隣接文間で、決定木は隣接関係ありのクラスに分類するが、人間の被験者が三者一致で話題境界としている例である。

決定木学習では、文頭提題で提示された主題が、直前文の内容と強い類似性や関連性を示す場合、直接的な部分内容説明や補足をしているとして隣接関係ありに分類する決定木が統計的に得られている。しかし、この例の場合、2 文目の下線部分が文頭提題部分で、話題「多国間投資協定」が文頭に現れ、そこでの文字が直前の文において相当数使われており字面類似度が高くなっている。このため、決定木は文頭で新しく提示された話題が直前文との関連が高く、隣接文間で話題が継続しているものと判断するが、人間の被験者は三者一致で話題境界と判断しているため、分類誤りとなる。

このような例を決定木によってうまく分類するためには、2 文間の記述内容の比較が字面の類似度だけでは不十分であり、少なくとも、下線部の「多国間投資協定」が直前文での記述内容に対してどのような位置付けになるかを同定する必要がある。今回、NE 属性を隣接文間の記述内容を特徴付ける属性として利用したが、固有名詞のうち人名、地名、組織名のみしか扱っていないため、このような事例を扱うにはなんらかの拡張が必要である。

固有名詞や字面の類似度が記述内容属性としてうまく機能していても、決定木により人間が分類したクラスに分類できなかった例を例 2 に示す。この隣接文間は形式段落境界ではなく、決定木は話題境界と判断するが、人間は三者一致で隣接関係ありと判断する例である。この隣接文における第 2 文は、文頭提題部分「運輸省は」が字面の類似度を用いても、NE 属性を用いても、新規の話題「運輸省」(組織名として認識) が文頭に出現したと分析され、2 文間に同一対象を参照する表現がないことも分析される。その結果、決定木は隣接 2 文の間で記述内容の変化があるとして話題境界と判断するが、人間の被験者は三者一致で隣接関係があるとしてしまう分類誤り例である。

このような例を計算機によって判断させるためには、文頭提題部と文末表現を単に別々の属性と値の組で表すだけでなく、話題導入のタイプを分類する必要があると考えられる。また、例 1、例 2 に共通する課題として、隣接 2 文のみの言語情報だけではなく、その隣接文間を境界とする文のまとまり同士の記述内容を比較する必要があると思われる。

5 まとめ

今回の実験では、試作したテキスト構造解析タグ付け体系を用いて被験者による新聞報道記事のテキスト構造解析実験を行ったところ、局所的な関係については比較的一致することがみられた。また、人間による話題境界の判断の傾向を言語情

例 1: 決定木は隣接関係ありに分類するが、人間は話題境界と判断した例

企業の海外移転や対外投資に伴って国内空洞化が進むなかで、地域共同体内の投資優遇策など地域主義を監視・けん制し、各国の収益を守る狙いもある。

多国間投資協定については、先進国で構成する経済協力開発機構（ＯＥＣＤ）が十八―二十一日に開く「資本移動貿易外取引」「多国籍企業」の両委員会の合同委員会で骨格を固める。（日本経済 95 年 4 月 16 日）

例 2: 決定木は話題境界に分類するが、人間は隣接関係ありと判断した例

釜山港や台湾・高雄港が新たな貨物船基地に成長しつつあるほか、日本向け航路がシンガポールや香港からの支線の地位に転落する兆候も出ている。

運輸省は日曜荷役再開を「港湾の国際競争力強化につながる」（海上交通局）とみている。（日本経済 95 年 6 月 5 日）

報から検討したところ、形式段落が最も影響を及ぼしていることがわかったが、それ以外の言語情報を用いるだけでもある程度人間のテキスト解析の傾向を把握できることがわかった。具体的には、いくつかの言語情報を整理して決定木学習を行って得られた決定木を用いて正解率を評価した値が、形式段落属性を用いた場合が 76.0% であり、形式段落属性を用いずにその他の言語情報のみを属性として利用した場合は 73.9% であった。

本実験に関連する先行研究としては、談話セグメンテーションを行う研究 [12, 3] などがあげられる。しかし、本研究で分類しているのは隣接文間のみを対象とした局所的な話題境界であるため、正解率によって単純な比較をすることは難しい。今後、そのような談話セグメンテーションとの融合を視座にタグ付け体系を改良し、新聞報道記事以外にも対象を広げ、テキスト構造の特質を明らかにしてゆくことが課題である。また、実験で得られたテキスト構造と人間が作成した要約との関係を比較検討することも行ってゆきたい。

謝辞

本実験の実験データとして新聞報道記事を使用させていただいた日本経済新聞社に心から感謝します。また、本実験のタグ付け実験に協力してくださった方々に、この場をお借りし、お礼申し上げます。

参考文献

[1] 奥村学、難波英嗣: テキスト自動要約に関する研究動向、 自然言語処理, Vol.6, No.6, pp.1-26,(1999).

[2] 比留間正樹、山下卓規、奈良雅雄、田村直良: 文章の構造化による修辞情報を利用した自動抄録と文章要約、 自然言語処理, Vol.6, No.6, pp.113-129,(1999).

[3] 望月源、本田岳夫、奥村学: 複数の表装的手がかりを統合したテキストセグメンテーション、 自然言語処理, Vol.6, No.3, pp.43-58,(1999).

[4] 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).

[5] J. Carletta, A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson: The Reliability of a Dialogue Structure Coding Scheme, *Computational Linguistics, Vol.23, No.1,pp. 13-31*,(1997).

[6] K. Ono, K. Sumita, S. Miike: Abstract Generation Based on Rhetorical Structure Extraction, *COLING-94*, Vol.1,(1994).

[7] W. C. Mann and S. A. Thompson: *Rhetorical Structure Theory: A Theory of Text Organization*, Technical Report ISI/RS-87-190, ISI Reprint Series,(1997).

[8] Daniel Marcu: Improving Summarization through Rhetorical Parsing Tuning, *Proceedings of the Sixth Workshop on Very Large Corpora* pp.206-215,(1998).

[9] Daniel Marcu: To Build Text Summaries of High Quality, Nuclearity Is Not Sufficient, *Intelligent Text Summarization *Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06*, pp.1-8,(1998)

[10] J. D. Moore and M. E. Pollack: A Problem for RST: The Need for Multi-Level Discourse Analysis, *Computational Linguistics, Vol.18, No.4*, pp.537-544,(1992).

[11] M. Moser and J. D. Moore: Toward a Synthesis of Two Accounts of Discourse Structure, *Computational Linguistics, Vol.22, No.3, pp.409-419*,(1996).

[12] R. J. Passonneau and D. J. Litman: Discourse Segmentation by Human and Automated Means, *Computational Linguistics, Vol.23, No.1*,pp.103-139,(1997).

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