修辞構造のアノテーションに基づく要約生成*

綾 聡平† 橋田 浩一‡ 石塚 満§

東京大学大学院情報理工学系研究科 †§ 産業技術総合研究所サイバーアシスト研究センター ‡
科学技術振興事業団 CREST‡

1 はじめに

新型肺炎 SARS や牛海綿状脳症 BSE 等，様々な事件に関連するニュースが，連日に渡って報道されている．しかし，そういった事件に関する顛末を知るべく，その全てに目を通すのは極めて大きな負担である．本稿では，このようなある事件に関する一連の記事に対して要約を施し，ユーザに簡潔にって維持するシステムを提案する．

自動要約の上で，最も大きな問題のひとつとして挙げられるのが一貫性，つまり文章全体としてのまとまりを維持することである．この問題を解決すべく考えられたのが，文間のつながりに着目した修辞構造理論（Rhetorical Structure Theory，以下 RST）である．本稿では，この RST を拡張したものを用い，複数文書を対象とした要約を行うことを目的とする．

また，2 文間の関係は，接続詞等の手がかり語等に基づいてある程度自動化する手法も考えられているが，高精度の自然言語処理を完全自動化することは難しい．そこで，本研究では GDA と呼ばれるアノテーションを人手でつけた文書を要約対象とする．

2 提案手法の概要

図 1 に提案手法のシステムの概要を載せる．この詳細について，以下に概説する．

2.1 大域文書修飾 GDA

まず，要約対象記事にアノテーションとして用いている GDA について簡単に説明する．

意味や常識，概念等をコンピュータに理解させることは困難であり，それ故に自然言語処理を高精度に自動化することに成功した例は少ない．そこで，このような状況を改善すべく考えられたのが GDA（Global Document Annotation）である．紙面の関係上，詳細については省略[1]するが，XML タグセットにより，既存の自然言語解析ツールに加え人手で文書中の照応・共参照，修辞構造等の情報を予め明示的に示しておくことにより，計算機による自然言語処理の精度を高めることを目的としている．

*Summarization of Multiple Documents with Rhetorical Annotation
†Soiki AYA
‡Hasida KÔITI
§Mitsuru ISHIZUKA
[1]詳細は http://www.i-connect.org/gda/

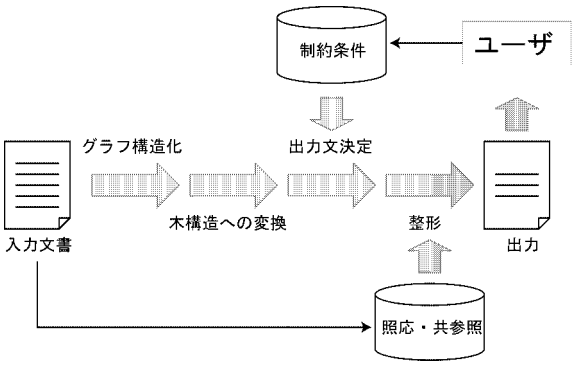

図 1 システム概観

```
<su id="node01" background="node02">
  <orgname id="nihonTV">日本テレビ放送網（本社・東京都港区）</orgname>の男性社員プロデューサー（４１）が、視聴率調査会社「ビデオリサーチ」の調査対象世帯に現金などを渡して自分が制作した番組を見るように依頼していたと<datep value="2003-10-24">２４日</datep>、<orgname eq="nihonTV">日テレ</orgname>が発表した。
</su>
<su id="node02" cause="node03">…</su>
```

図 2 GDA のアノテーション例

GDA のサンプルを，図 2 に示す．この例に於いて，文を示す su 要素中に記述されている属性，“background”，“cause” が修辞関係を記している．

2.2 グラフ構造化

GDA により与えられた修辞構造を元に，文書たちをグラフ構造とする．この構造において，文がノードに，文間の修辞関係がリンクに相当する．利用する修辞関係は，これまでの RST を参考に，53 種定義した．

従来の RST では，単一文書を対象とし，ここに含まれるテキストセグメントたちに含まれる修辞構造を木構造として捉えていた．しかし，木構造として無理に捉えることにより，実際にはあるはずの修辞関係が，排除される可能性がある．また，本研究では対象は複数文書であり，全体として木構造を成すと捉えるのは無理がある上，全体的な構造を考えながらアノテーションするのは非常に負担が大きい．さらに，出力文の中心トピックを，ユーザの意図により変更し，即ち書き手の意図に依らない出力を再構成するシステムを目指している為，文書に対する核と衛星といった重要度の差を認めるのは適当で

ない．

そこで，今回，木構造とならない修辞構造を認め，さらに其々の関係を対象リンクとして捉えるよう，拡張を図った．

2.3 木構造への変換

第2.2節で木構造とならない修辞構造を認めたものの，最終的な出力形態としては，ある文を中心とする要約を作成すべく，グラフ構造を木構造に変換する．そこで，グラフに於いてサイクルがあった場合に，ノードを分割することを考える．なぜリンクではなくノードを分割するかというと，リンクを切ると，あったはずのノード間（文間）の関係が消滅してしまい，第2.2節で図った，修辞構造拡張の意図に反するからである．

尚，サイクルに於いて分割するノードは，持っているリンクの種類と，持っているリンクの数により決定する．

2.4 出力文決定

まず，要約出力の中心となるノード（文）を決定する．この中心ノードはユーザの入力により決定することもできるが，このような入力がなかった場合には，活性拡散に基づく重み付けにより，これを決定する．

活性拡散は，以下の式で与えられ，この計算を活性値が収束するまで行う．

$$\mathbf{A}(t) = \{(1-\rho)\mathbf{I} + \rho\mathbf{R}\}\,\mathbf{A}(t-1)$$

ここで ρ は減衰定数，$\mathbf{A}(t)$ は活性値，$\mathbf{R}$ はネットワークの接続行列である．この値を元に，子孫ノードの活性値の和を其々のノードの重要度とする．

このようにして決定した重要度に基づき，中心ノードから順に，字数制限に達するまで，重要度の高いノードを展開していく．

2.5 出力順決定

第2.4節で出力することが決定したノードたちに対し，中心ノードから順に以下の処理を再帰的に行うことで，出力順を決定する．

まず，あるノード $\mathbf{A}$ からリンクが張られたノード $\mathbf{B}$ を $\mathbf{A}$ の前方に置くか，後方に置くかは，ユーザに依る制約，修辞関係の種類，日付の前後等に基づいて決定される．

さらに，方向が決定したノードたちに対し，$\mathbf{A}$ からリンクが張られたノード $\mathbf{B}$ とノード $\mathbf{C}$（両方前方，もしくは両方後方とする）のどちらをノード $\mathbf{A}$ の近く（遠く）に置くかについて決定が必要である．この際には，ユーザによる制約，子孫ノード数，日付の前後等が考慮される．

広東省広州市で肺炎と診断されて入院中の男性患者（３２）に、新型肺炎（ＳＡＲＳ）の疑いがある。2003年12月20日に中山大学第１付属病院で肺炎と診断され入院、隔離された。患者は広州市内に住むフリーのテレビ番組制作者。同日にＳＡＲＳの疑いがあると診断された同省広州市の男性（３２）の遺伝子配列について、野生動物のハクビシンのＳＡＲＳコロナウイルスとほぼ一致した。中国広東省疾病予防センターは「人類のＳＡＲＳの感染源はハクビシンである可能性が高まった」としている。中国本土では同年、広東省から広がったＳＡＲＳで５３００人余が感染し、約３５０人が死亡した。感染情報の隠蔽（いんぺい）や、病院間の連絡が機能せず感染拡大を招いた教訓から、９月中旬から連日、国内の感染状況を公表するとともに、関係部門が情報を共有するネットワークづくりを進めてきた。中国の衛生省は同年春、ＳＡＲＳ被害について実際より小さく発表し、「うその報告だ」との元医師の内部告発や国際的批判を受けた。その結果、多数の犠牲者を出す一方、国際的にも中国のイメージはダウンした。

図3　出力結果例

2.6 整形

以上のように決定した出力文に対し，照応・共参照，日付の出力等を読みやすく変換したり，修辞関係に基づいて接続詞を填補してやることにより，より読みやすい出力文へと整形する．

3 出力結果

図3は，2003年12月に中国で新型肺炎SARS患者の疑いに関する，アサヒ・コム[2]の3つの記事（全約1800文字）を500文字の制限の元で要約した出力結果である．この出力の中心となっている文は，第1文である．

この出力に於いて，初めの3文が1つ目の記事から，4，5文が3つ目の記事から，第6文以降が2つ目の記事から選択されたものである．記事集合全体から，満遍なく文を選択できていることがわかる．

また，中心文や文の順序（前方，後方等含む）は，ユーザの意図を制約ファイルに書き込むことで，出力結果に反映することが可能である．

4 結論と今後の課題

本稿では，修辞構造のアノテーションに基づいた要約手法を提案した．修辞構造を利用することにより，一貫性に優れた要約が出力できることがわかった．さらに，照応・共参照や日付等のアノテーション情報を利用することで，読みやすい出力結果を得ることも可能である．

今後は，自動解析した修辞構造に基づく要約作成や，複文等に対する，より小さな単位での解析を行えるようにするべく，システムを改良する予定である．

[2] http://www.asahi.com/