談話文からの命題-様相の抽出システム

永野 圭一郎† 辻井 潤一‡ 鳥澤 健太郎‡*
†東京大学 理学部 情報科学科
‡東京大学大学院 理学系研究科 情報科学専攻 * さきがけ研究21
{gano, tsujii, torisawa}@is.s.u-tokyo.ac.jp

1 はじめに

文書は、ある目的にしたがって筆者が各種の情報を組織化して提示したものである。この文書という筆者中心の情報の組織化を解きほぐし、情報の消費者にとって有効な情報を取り出す研究は、情報抽出や Passage 検索、あるいは、 TREC の Q/A タスクとして研究されている。

文書中の情報を縮約する自動抄録の研究は、情報消費者がそこから関連性を判断できる示唆的抄録 (Indicative Abstract) を作る研究と消費者が必要とする情報を保った抄録 (Informative Abstract) を作る研究に分かれ、後者は、情報抽出・Passage 検索の研究へと接近してくる。また、後者では、抄録対象を一つの文書に限定する必要はなく、多テキストからの抄録の可能性も提案されている。

一方、文書を口語体や対話の形態に変換することで、その理解容易性が格段に向上するとの提案もある。我々のグループでは、これらの研究が目指す文書情報の形態変換処理の典型として、2つの独立した文書から対話形態の抄録を作る研究を行っている。

本報告では、この研究構想の概要、および、その第一歩として、文書情報から命題と様相の情報を分離・抽出する研究について述べる。

2 問題の設定

従来の情報抽出の研究では、経済情報、製品情報、人事異動などの客観的な事実情報を文書中から抽出することが中心であった。しかし、新聞社説をその典型とする論説記事では、筆者の伝達する情報は、筆者の主張である。この中心となる主張を読者に伝達するために、筆者は客観的な事実、それに対する筆者の評価、判断にもとづく要求、といったものを組み合わせて一種の論理的な筋道を構成する。したがって、論説文の抄録作成、とくに、情報を保ったままでの抄録作成では、この論理展開の骨組みを把握することが重要であり、このために W.Mann らの修辞構造の理論を適用する研究が行われてきた [1]。

羽田氏: 不正な株取引で利益を得ていたという自民党の新井将敬代議士の疑惑は、現実のものとなり、逮捕許諾請求要求書が提出された。新井氏は自らの進退を明らかにし、職を辞するべきだ。

首相: 内閣は東京地裁から新井氏の逮捕許諾請求要求書を受理し、持ち回り閣議を経て、衆院に逮捕許諾を請求した。所定の手続きを行い、速やかに逮捕許諾の判断がなされることを期待する。自民党は新井氏に事実上の離党勧告を行い、党としての考えは伝えているところだ。職を辞すかどうかは新井氏個人の判断を待ちたいと思う。事実は司法の手によって明らかにされると考えている。

図 1: 質疑応答の実例 (日経新聞 98.02.19 朝刊)

一方、筆者の判断、要求は筆者固有のものであり、同一の事態に対して全く正反対の主張を展開することも可能である。同一の事態に対する複数の論説文を対象にし、その論点の差を明確にした抄録の作成は、多テキストからの抄録作成の課題として興味ある問題となる。

我々は、この種の問題意識から、政府と野党の間の国会答弁書から、政府と野党との間の仮想の質疑応答対を作成することを目標とした研究を開始している (図1)。具体的には、野党の質問書、それに対する政府の答弁書を、文書として衆議院ホームページ [2] から入手し、この2つの文書から短い応答対の系列を作成することである。実際、新聞記事では記者が人手で整理した応答対が載せられているため、これを正解例として参照することもできる。

同種の問題は、環境アセスメントに対して寄せられた文書から、中心となる論争点を整理するドイツ GMD のシステム [3] があるが、このシステムは、人手による整理が前提となっており、言語処理技術は使われていない。

3 システムの全体構成

単一文書の抄録作成、特に、示唆的な抄録の作成では、文書中で重要な役割を果たす文の認識が中核となる。いったん、重要な文が認識できると、その文を理解するのに必要な範囲の認定をし、その範囲内の文を表層的に短縮することで抄録が作成できる。また、重要な文の認識も、

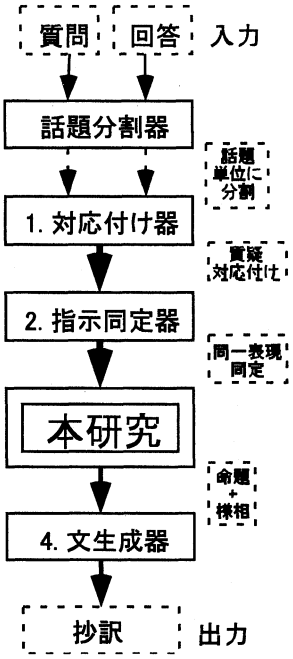

図 2: システムの全体構成

tf×idf などの統計的な尺度 [4, 5, 6, 7] で可能となるために、言語的な構造の処理によらなくても、比較的良い抄録作成が可能である。

しかし、前節で述べたような多テキストからの抄録作成では、筆者間の認識の差を明示的に把握し、その差を明示する必要があるため、表層的な抄録作成手法だけでは対処できず、分かりやすい応答対を作成するためには、構造的な言語処理の手法を必要とする。

システムは、図 2 に示すように次の 4 つの処理段階から構成される。

1. 2 つの文書中で類似の内容を記載した部分の対応をとるアライメント処理
2. アライメントされた部分中で、同一指示の関係にある表現の同定処理
3. 同一指示についての筆者の認識・判断の差を認定する処理
4. 応答対（抄録）の生成処理

仲尾らは、文書集合中の語彙の類似をもとに、まず単一文書内での階層的なパラグラフ認定を行い [8]、次に野党・政府の文書のパラグラフ類似度を計算することで、段階 1 の処理が可能であることを示している [9]。本研究では、段階 1 については、この手法を用いることを前提としている。また、段階 2 については、従来の照応処理の複数文書間への拡張で行えることを前提とする。

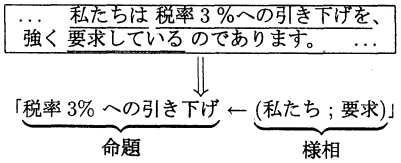

図 3: 命題と様相

賛成	承知しています/よくわかります
反対	疑問であります/反対をいたします/賛成しがたい
確信	確信をいたします/重要であります/当然であります/言をまちません
推測	考えます/思います

表 1: 様相を示す語句の例

したがって、本稿では、ある程度短いパラグラフ中の文から、客観的な事態（**命題**と呼ぶ）とそれに対する判断・要求といった**様相 (モダリティ)** の部分を分離すること、また、その**判断の主体**を認識する処理に関して以下に詳述する。なお、文書中の文は、我々がすでに開発した日本語係り受け解析プログラム [10, 11] によって解析されている。

4 命題と様相

我々は、談話文の主張単位ごとに、客観的事態を述べる部分と、それに対して話者が主観を述べる部分とがあるとする。これらを、それぞれ「**命題 (propositional content)**」「**様相 (modality)**」と名付ける。

例を 図 3 に示す。ここで “税率 3% への引き下げ” は事実を示すもので、それに対する様相をあらわす語句 “要求する” のヲ格に入っている。このような様相を示す語句にどのような種類があるかを、“賛成/反対” “確信/推測” の 4 種類について 表 1 に例示する。

このような様相を示す語句は、文章の表面上にあらわれていないものもありうる[1]。しかし今回は、対象とした国会答弁質疑応答の性格を鑑み、文面上に明示されている様相のみを取り扱う。我々が提案する自動抄録は、まず原文から主張を命題と様相の形で抜き出し、そこから文生成を行うことで要約としようというものである (図 4)。

[1]発話行為 (Speech Act [12])。例えば、発話「暑いと思いませんか?」は、「窓を開けて欲しい」という illocutionary act を示すことがある。ここには、我々が議論している「要求」の様相が含まれている。

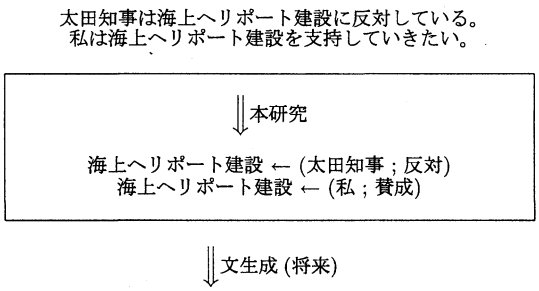

図 4: 抄録過程の例

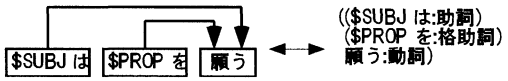

図 5: ルールの例

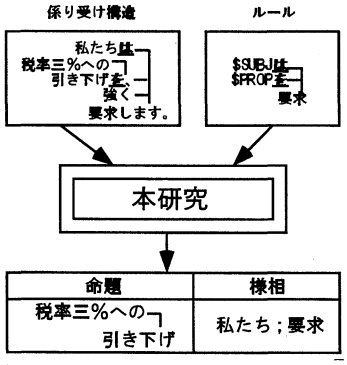

図 6: ルール適用

5 抽出システム

このような命題と様相の抽出のために、我々は係り受け構造に対するルールを記述するという方法を提案する。ルールは様相を示す単語をベースにした部分係り受け構造の形とし、様相の主体や命題を要求する部分に変数を書き込んでおく。図 5 に例を示す。このような構造のルールを用いて抽出を行うことは、変数つきの部分係り受け構造と原文の係り受け構造との間で、論理プログラミング言語でいう単一化計算 (unification) の操作を行うことに相当する。ルール適用過程の概略を 図 6 に示す。

正規表現によるルール記述と比較した時の、この手法の利点は以下の 3 点である。

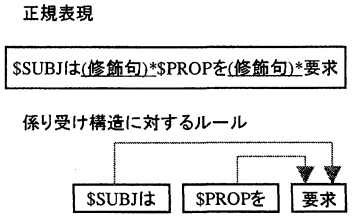

図 7: 修飾語句を考慮した記述

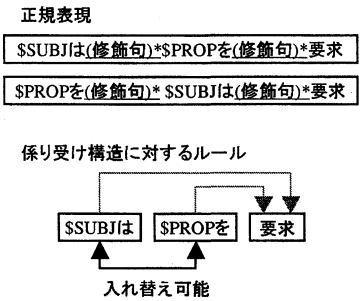

図 8: 語順入れ替えを考慮した記述

1. 修飾語句をまたがった抽出
   正規表現でルールを記述する場合、途中の修飾語句の読み飛ばしを意識しなければならない。係り受け構造のルールにおいては、修飾語句は単に省略すればよい (図 7) 。
2. 語順の入れ替えに対応した抽出
   日本語は語順が比較的自由であるという特徴を持っている。正規表現であれば、語順の全通りを考慮した複数のルールを記述して対処しなければならない。部分係り受け構造の形をしたルールであれば、システム側で適用時に語順を入れ替えることが可能である (図 8) 。
3. 深い構造からの抽出
   埋め込み文からの抽出を行う場合、正規表現によるパターンでは、図 9 のような場合に 2 つの助詞「は」の係り先の曖昧さを解消できない。係り受け構造であれば、修飾関係から、「要求」の主語を直ちに決定することができる。

6 実験結果

上記のような部分係り受け構造単一化のシステムを実装し、様相を示す用言「要求」「願う」について 4 つずつのルールを書き、それらの単語を含む文・全 40 例よ

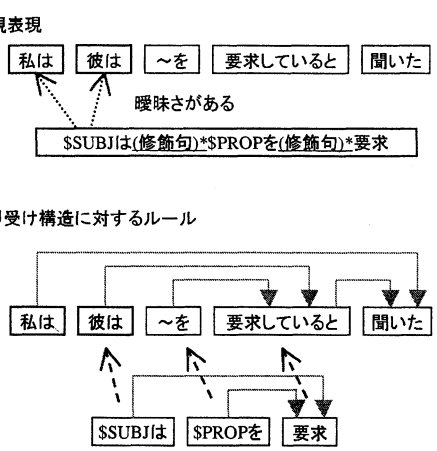

図 9: 深い構造からの抽出

り 50% の文から正しい命題-様相を抽出することに成功した。

失敗した例としては以下のようなものが挙げられる。

- 目的格「を」が省略されていたもの (8 例)
- 用言ではなく名詞の用法であったもの (7 例)
  ex. 「全野党の修正要求に」
  → 名詞用法のためのルールを書けば対処可能
- 係り受け解析器の間違い (3 例)
- 助詞「は」の曖昧さによる間違い (2 例)
  ex. 「米朝間交渉では食料を要求」

ただし、成功例として挙げた中には、(ゼロ代名詞によって) 助詞「は」「が」で明記された主語がないため、様相の主体が明確に得られないものが 15 例見られた。様相は判断の主体を要求するが、本システムは主語省略に対処することができない。これは、[13, 14] の研究を応用し、省略補完の前処理に通すことで対処できると考えている。

7 総括、および今後の方向

本報告では、多テキストからの、情報を保った抄録作成について考察した。その一例として国会議事録質疑応答を取り上げ、そこから主張を**命題**と**様相**とに分けて抽出する方法を論じた。その方法とは、部分係り受け構造の形をしたルールを用いるというものである。システムを実装し、実験を行い、今後の方向性を明らかにした。

今後は、何にも増してルールの数を増やし coverage を上げるのが先決である。が、その際にも役立つ手法として、**ルールの変形導出機構**を考察している。これは、変形表現による係り受け構造の変化をシステム側に保持しておき、それに従ってルールを適宜変形して適用するというものである。これによって、より少ないルール数で命題・様相表現がカバーできるようになることが期待できる。

参考文献

[1] Jun'ichi FUKUMOTO and Jun'ichi TSUJII. Breaking down rhetorical relations for the purpose of analising discourse structures. In *Proc. of Coling*, pp. 1177–1183, Kyoto, Japan, 1994.

[2] 衆議院議事録ホームページ. http://www.shugiin.go.jp/.

[3] Thomas Gordon, Nikos Karacapilidis, and Hans Voss. Zeno: a mediation system for spatial planning GMD-FIT, 1996. Sankt Augustin, Germany (http://orgwis.gmd.de/projects/W4G/proc.html).

[4] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, Vol. 24, No. 5, pp. 513–523, 1988.

[5] 大澤幸生, ネルス E. ベンソン, 谷内田正彦. KeyGraph: 語の共起グラフの分割・統合によるキーワード抽出. 電子情報通信学会論文誌 D-I Vol.J82-D-I No.2, pp. 391–400, Feb 1999.

[6] Yukio Ohsawa, Nels E. Benson, and Masahiko Yachida. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor.

[7] Regina Barzilay and Michael Elhadad. Using Lexical Chains for text summarization. In Inderjeet Mani and T. Maybury, editors, *Advances in Automatic Text Summarization*, chapter 10, pp. 111–122. The MIT Press, 1999.

[8] 仲尾由雄. 語彙的結束性に基づく話題の階層構成の認定. 自然言語処理, Vol. 6, No. 6, pp. 83–112, July 1999.

[9] 仲尾由雄. 話題の階層構成に基づく関連談話の対応付け. 情処研報 FI-60-4, Sep 2000.

[10] 金山博, 鳥澤健太郎, 光石豊, 辻井潤一. 3 つ組・4 つ組モデルによる日本語係り受け解析. 言語処理学会第 6 回年次大会, pp. 487–490, 2000.

[11] KANAYAMA Hiroshi, TORISAWA Kentaro, MITSUISHI Yutaka, and TSUJII Jun'ichi. A Statistical Japanese Dependency Analysis Model with Choice Restricted to at Most Three Modification Candidates. *Journal of Natural Language Processing in Japan*, Vol. 7, No. 5, 2000.

[12] J. L. Austin. How to do Things with Words. *Oxford*, 1962.

[13] Barbara Grosz, Aravind Joshi, and Scott Weinstein. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, Vol. 2, No. 21, pp. 203–225, June 1995.

[14] Kameyama Megumi. *Zero anaphora: the case of Japanese.* PhD thesis, Stanford University, Linguistics Department, 1985.