係り受け解析による独話構造解析

柏岡 秀紀
ATR 音声言語コミュニケーション研究所
hideki.kashioka@atr.jp

1 はじめに

講演会や学会発表などの独話は貴重な知的財産であり、マルチメディアデジタルデータとして蓄積することにより、膨大な知的資源となる可能性を秘めている。これらの独話データを知的資源として効率的に活用するには、膨大な知的資源への効率的なアクセスの実現が望まれる。これを実現するためには、独話の書き起こしテキスト、およびそのテキストへの様々な付加情報を含むタグつきデータの構築が必要である。各独話の談話構造の情報は、有効な付加情報の一つである。談話構造が明確になると、要約の作成や高度な検索が可能となり、データへのアクセス効率が高まるためである。談話構造解析は、文と文との間の関係を把握することが目的であり、照応解析や省略語の解析、話題の解析の結果を利用することにより把握する手法（逆に、談話構造を利用した照応解析や省略語の解析を行う手法もある）などが研究されている。

文章として書かれたテキストの談話構造は、筆者により精巧に構築されたものが多く、その構成要素である文も推敲されており、適切な長さの文から構成されていることが多い。しかし、独話においては、原稿を読み上げている形式の独話であっても、書き起こしたときの文の長さや文の接続は推敲された文章とは異なり、極端に短いものから長いものまで様々なものが含まれている。また、文と文との関係で、適切とはいえない接続のものもある。文の長さが多様になってしまうのは、講演者が聴衆の反応を見ながら発話するためでもある。例えば、

- 「新しいゲーム機が発売されました。」
- 「非常に若者の人気を集めています。」

という文からなるテキストは、実際に発話されるときには、以下のような発話になっても不思議ではない。

- 「新しいゲーム機が発売され、若者の人気を集めています」

また、この発話は、**(新しいゲーム機が発売され)** と **(若者の人気を集めています)** という二つの**節**の関係となっており、(発売され) が (集めています) に係るという係り受け構造で示すことができ、既存の係り受け解析により解析できる。つまり、「XXX されました」＋「YYY(し) ています」という文の列を「XXX され、YYY(し) ています」に書き換えることで、理論的には、従来の係り受け解析の枠組みにより複数の文間の解析が可能になると思われる。ところが、従来、文長が長いほど係り受け候補が増大し、文の係り受け解析の実行は困難になり、解析精度も低くなっていた。大野ら [3] は、まず文を節に分解し、節の情報を利用して長文の解析を効率的に実現している。

本稿では、文末を書き換えることにより、本来、途切れている部分を節として次の文につなげ、大野らの係り受け解析手法を適用し独話全体の構造を構築する手法を提案する。また、提案手法の問題点等に考察を加え、独話の談話構造解析について検討する。

対象とした独話データは、NHK の解説番組「あすを読む」250 番組である。

2 節情報を利用した係り受け解析

本論文で対象としているデータは、講演などの独話データである。独話は一人の発話者が連続して発話しているものであり、「文」の境界を明確に判定することは困難である。また、書き起こし作業者による「文」の境界をみても、比較的長い文が多く見られ、その判定にもある程度の揺れがみられる。

長文においては、係り受けの候補が多くなり、解析精度が悪くなると共に、解析速度も遅くなる傾向が見られる。しかしながら、節に着目すると、節境界単位の内部では大半が閉じた係り受け構造を持つと報告されている [2]。大野らは、この特徴を生かし係り受け解析を 2 段階で処理する手法を提案している [3]。図 1 に

解析処理の流れを示す[1]。この図 1 に示されているように、節内部の解析が先に行われるために、節末の文節の係り先の候補は、非交差の原理によりかなり絞りこむことができる。また、節分割には、CBAP[1] を利用している。

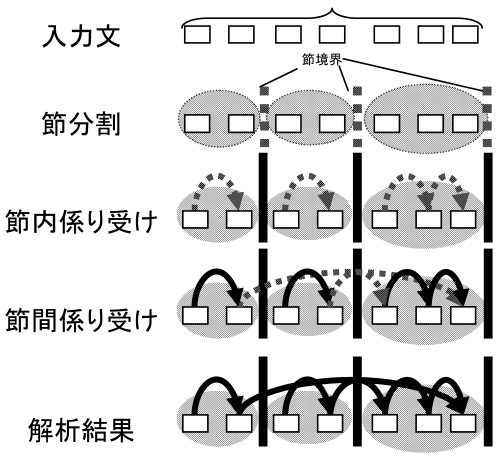

図 1: 係り受け解析の流れ

この係り受け解析において、2 段目に行っている節末の文節がどこに係るかという解析は、節と節との関係を示していると捉えることができる。そこで、主節の最終文節に対しても、係り受け関係を考えることができれば、文章全体での構造を把握することができると考えた。

3 文末の修正

前節で示したように、主節の最終文節に対して、従来の節末の文節の係り受け解析が適用できれば、文章全体での構造を把握することができると思われる。一方で、独話に含まれる発話は、「はじめに」でも述べたように、文末が明確ではなく、文末を節に言い換え可能と考えられる。

「あすを読む」250 番組には、15,313 個の文末が含まれている。その中で頻繁に現れた文末表現 (文末からの 4 形態素の連鎖)、上位 10 個の表現を表 1 に示す。

また、これらの文末表現に対して、例えば、表 2 に示すような単純な文末の書き換えが考えられる。このような書き換えを行った実際の発話の例を以下に示す[2]。

[1]図中の四角は、文節に相当する。

[2]ただし、このような書き換えは、常にうまくいくとは限らない。現在、書き換え規則の構築を行いつつ精度を検討しているところである。

表 1: 上位 10 個の文末表現

頻度	文末表現
865	と * 思い ます
444	し て い ます
219	なっ て い ます
218	れ て い ます
198	いう * こと です
183	て き まし た
154	し て おり ます
153	ない でしょ う か
150	の でしょ う か
134	ませ ん でし た

文末表現: ～が出るほどの接戦になっております。さて今回のアメリカ大統領選挙を見て、～
書き換え後: ～が出るほどの接戦になっておりますが、(さて) 今回のアメリカ大統領選挙を見て、～

文末表現: ～について考えてまいります。事件 5 つあります。～
書き換え後: ～について考えてまいりますが、事件 5 つあります。～

表 2: 文末を節に書き換える単純な規則

文末表現	節末表現
思います。	思いますが、
XX したいと思います。	XX し
XX しています。しかし	XX していますが、
XX しています。	XX しており、
XX になっています。	XX になっており、
XX(して) おります。さて	XX(して) おりますが、
XX します。	XX しますと、
XX(し) てまいります。	XX(し) てまいりますが、
XX がある。	XX があり、

実際に、文末を一つ書き換えて見た場合、それほど違和感なく読むことができることが多い。しかしながら、その連鎖を繰り返すと違和感を感じてしまう。これは、あまりにも長い文になってしまっていることと、微妙な接続関係のずれが生じているために感じる違和

感だと思われる。

このような書き換えを行うことにより、主節の最終文節の係り先の解析を従来の節末文節の係り受け解析により実現することができる[3]。

4 独話の構造

本節では、提案する独話の構造解析手法について説明する。処理の流れは、以下のようになる。

1. 入力されたテキストに対して節分割処理を行う[4]。
2. 最初の節から順に、節内部の係り受け解析を行う[5]。
3. 節内部の解析後、それ以前の節に含まれる節末文節の係り受け解析を行う。
4. 文末であれば、書き換えを行い、節と同様の処理を行う。

多くの場合、ある節の節末の文節は次の節の内部に含まれる文節にかかる。そのため、節の間の依存構造は、文節間の係り受けを下に見ると、図 2 のような単調な連鎖構造となることが多い。

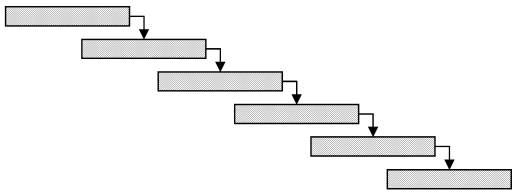

図 2: 単調な係り受けによる独話の構造

別の見方をすれば、発話順序を示しているだけとも考えられ、この構造だけでは、あまり有益な情報にはなっていないかもしれない。そこで、このそれぞれの関係がどのような関係であるのかが、より重要な要因となる。現在、この関係の分類を行っている。詳細は次節の考察で述べる。

基本的な係り受け解析に利用する情報は、節内部の解析に対しては従来から報告されている係り受け解析が利用している情報と違いはない。2 段目に行う節末の文節の係り受け解析を行う場合、節内部と同様の知識で可能な部分もあるが、異なった知識を利用すべきと思われる部分がある。利用すべき知識の異なりは、南 [4] による従属節の形態および従属度の対応付けに対応するものと思われる。この従属度は、ある程度、節ラベルに対応しており、節ラベルに応じて、知識を使い分けることで、緻密な解析が可能になると思われる。

[3]書き換え方によって係り受け解析結果が異なる可能性もある。可能な書き換え方とそれにより生じる係り受け構造の差異を確認する必要がある。

[4]CBAP を利用

[5]節内部の係り受け解析は、既存の係り受け解析手法を利用することが可能であり、特殊な手法に特化しているわけではない。

5 考察

現在、係り受け構造から見られる節と節との関係の洗い出しを人手により行っている (図 3 にその関係の一部を示す。順序は、人手によるラベル付けにおいて多くつけられたラベルの順である。)。

接続関係、前提の関係、詳細説明、話題転換、因果関係、具体化、問題提起、並列、反対・対立、反復、など

図 3: 節間の関係

これらの関係を階層化し洗練することで、係り受け構造を構成する表現から関係を推定できれば、従来の「起承転結」のような構造や話題の推移から構成される構造などの談話構造を見出すことができる。

また、従来の係り受け解析手法を素直に適用できるように文末の書き換えを行っているが、書き換えなくとも、同じ情報が、文末周辺の表現に含まれているため、文末表現と周りの属性から係り受け関係を学習することにより、解析できる。ただし、学習するための文末に関する係り受け構造を付与したデータが必要となる。

6 まとめ

本稿では、節分割を行い、節内部の係り受け解析と節間の係り受け解析を 2 段階で処理する手法を応用し、文末を書き換えることで、独話全体での係り受け構造を解析する手法を提案した。節内部と節間の解析を 2 段階で行うことにより、係り受け候補を絞ることができ、文末を書き換えて接続しても解析が可能になっている。得られる構造は、いわゆる起承転結のような構造ではなく、どちらかといえば、単純な連鎖構造となっている。しかし、文末近辺に現れる接続詞や接続表現の情報や係り受け関係の節の関係を分析することにより、複数の節をまとめ上げ、いわゆる談話構造の把握

が可能になると考える。今後、係り受け構造により構築された構造から、いわゆる談話構造への変換を行い、要約等の処理での構造情報の利用価値について検討していく予定である。

謝辞

本研究は総務省戦略的情報通信研究開発推進制度における研究委託により実施したものである。

参考文献

[1] 丸山, 柏岡, 熊野, 田中. 日本語節境界検出プログラム cbap の開発と評価. 自然言語処理, Vol. 11, No. 3, pp. 39–68, 2004.

[2] 柏岡秀紀, 丸山岳彦, 田中英輝. 節境界と係り受け解析. 言語処理学会第 9 回年次大会発表論文集, pp. 117–120, 2003.

[3] 大野, 松原, 丸山, 柏岡, 田中, 稲垣. 節境界に基づく独話文係り受け解析の効率化. 情報処理学会研究会報告, pp. 213–220, 2004.

[4] 南. 現代日本語の構造. 大修館出版, 1974.