Vol.2013-NL-211 No.2
Vol.2013-SLP-96 No.2
2013/5/23

日本語書き言葉を対象とした談話単位分割基準の提案と自動分割の評価

宮原 聡[1,a)]　飯田 龍[1,b)]　徳永 健伸[1,c)]

概要：文を談話単位と呼ばれる基礎的な単位に分割する処理は談話関係解析などの前処理として必須である．ただし，談話単位間に論理的な談話関係を想定する場合には，談話単位に適切な粒度で命題が含まれる必要がある．これは，談話単位間で論理的な談話関係を想起する場合に，一つの談話単位に命題に相当する情報が含まれない場合には，関係を人手で付与する場合に解釈が困難になったり，また一つの談話単位に複数の命題が含まれている場合にはどちらの命題と関連させて関係を付与するのかわからなくなるという問題があるためである．本稿では談話単位の認定基準について議論し，談話単位アノテーションの仕様を設計し，日本語書き言葉均衡コーパス（BCCWJ）の一部に人手でアノテーションを行った．さらに，談話単位の境界にどのような特徴が現れるのかを人手で分析し，それらを手がかりとした自動分割の手法を提案する．この手法の有効性を調査するために BCCWJ にアノテーションした結果を利用した評価実験を行った結果について報告する．

1. はじめに

修辞構造理論（RST）[11] に代表される談話関係の理論に基づいて文章の談話の解析を行う際，文章を談話単位と呼ばれる単位に分割する処理は談話関係解析の前処理として必須となる．一般にこの単位は文より短いため，談話関係解析の観点から意味のある粒度で文を分割する必要がある．英語を対象にした RST に関するアノテーション [12] では，概ね節をこの談話単位とみなして分割を行っているが，日本語で同様に節を対象に談話単位分割を行うと命題を構成する内容が必要以上に細分割されてしまい，以降の談話関係解析の処理が難しくなる．例えば，例 (1) では並列節，連体修飾節を考慮して分割すると「国際会議で 2 件発表し」「論文誌が 1 件採録される」「ことが博士を取得するための必要条件となる」の 3 つに分割にすることができるが，このうち「ことが博士を取得するための必要条件となる」はこの節を修飾する要素が欠けているため，談話関係を判断する場合には情報が不足していることになる．

(1)　国際会議で 2 件発表し，論文誌が 1 件採録されることが博士号を取得するための必要条件となる．

また，「〜する前には」「〜する場合」などの表現をともなう場合，これらの表現の後で分割することが望ましいが，「前」や「場合」などが名詞に相当するため，品詞から単純に節を推定すると連体修飾関係にあると解析されてしまい，「前」や「場合」の前で分割されるという形態素の情報との互換性の問題も含まれる．

このような問題を解決するために，本研究では談話単位分割のための基準を規定し，その基準に基づいて日本語書き言葉均衡コーパス（BCCWJ）[7] の一部にアノテーションを行い，アノテーションされた談話単位の振舞いを学習することで自動的な談話単位分割を実現する．この談話単位分割の手法を評価するために，交差検定と新たな評価用データを用いた実験を行うことで，分割手法の有効性を調査する．本稿では，まず 2 節で本研究と関連する日本語の節境界認定や文分割の先行研究を紹介し，次に 3 節で解析対象とした文章とその文章に対して行った談話単位のアノテーション作業の仕様を説明する．4 節で談話単位分割に必要だと考えられる特徴について分析した結果について報告し，さらにその特徴に基づいた分割手法を 5 節で提案する．6 節で評価実験とその結果に対する考察を行い，最後に 7 節でまとめと今後の課題について述べる．

2. 先行研究

日本語を対象とした文分割処理について，さまざまな単位への分割が考えられる．例えば，武石ら [1] は読み手が理解しやすい文章を作ることを目的に，マニュアル文章中

[1] 東京工業大学
Tokyo Institute of Technology
[a)] miyahara.s.aa@m.titech.ac.jp
[b)] ryu-i@cl.cs.titech.ac.jp
[c)] take@cl.cs.titech.ac.jp

の複文を単文に分割する手法を提案している．彼らは，「ながら」や「つつ」などの接続表現と述語がともなうモダリティを手がかりに従属節の独立性を決定し，独立性が高い場合に分割している．例えば，例 (2) では「〜ので」と「〜けど」の 2 箇所で分割することが考えられるが，彼らの規則では逆接の「けど」の方が独立性が高いため，「けど」の後で分割する．

(2) [彼が勧めた〈ので〉$_B$ 行ってみた〈けど〉$_C$，][それほどでもなかったよ．][*1]

武石ら [1] の分割手法では形態素解析は行うが係り受け解析を利用しておらず，彼らも従来の構文解析技術と融合させることによって分割の精度が向上することについて言及している．そこで，本研究では係り受け解析の結果も利用して分割の性能の向上を図る．また，分割の粒度としては，本研究で扱う談話単位の粒度は彼らが分割したい文より細かいことに注意されたい．例えば，例 (2) では，彼らの手法では「〜ので」の後が分割対象となっていないが，「彼が勧めたので」と「行ってみたけど」の間には原因結果の関係があるため，談話関係解析を考えた場合，それぞれを談話単位とすべきである．このため，彼らが利用する接続表現に関する知見は利用するが，解くべき問題が異なるため，彼らが分類した接続表現のカテゴリを利用する代わりに接続表現を素性として利用することで，本研究で扱う分割の粒度に自動的に適合させる．

また，白井ら [2], [3] も同様に南らの接続表現の分類 [4] を拡張することにより，機械翻訳の前処理としての長文分割を行っている．この研究でも，接続表現ごとの従属節の独立性をあらかじめ決定しておき，独立性の高い箇所を分割することで，最終的な自動翻訳の品質が向上したことを報告している．

丸山ら [5] も同様に局所的な形態素列に着目して節境界の認定を行っている．彼らは話し言葉を対象にした節境界分割を局所的な形態素列の情報をもとに実現している．例えば，述語に後節して接続助詞「が」「けど」「けれども」などが出現する場合には，その箇所を強境界と認定し，その箇所で文を分割する手法を提案している．これの手法も上記の手法と同様に係り受け解析を想定せずに，節の分割を行い，その分割の強弱は出現する接続表現に基づいて分類を行っている．一方で本研究のように，文全体の構造を見て談話単位の境界を求める場合には，対象とする境界候補となる箇所が文のどのような位置に出現しているかを考える上で係り受け関係の情報は必須となる．この点について次節以降で説明する．

[*1] 文内にラベル付けられた B や C は彼らが南 [4] の接続表現の分類を参考に設計した接続表現の分類カテゴリを表す．詳細は [1] を参照されたい．

3. 人手による談話単位のアノテーション

本研究では日本語の書き言葉を対象にした談話単位分割を目的とする．このため，日本語書き言葉均衡コーパス（BCCWJ）[7] の一部を対象に談話単位のアノテーションを行う．この際，係り受け解析器 CaboCha [9] を用い，あらかじめ文節のチャンキングを行った結果を対象にアノテーションを行う．これは，アノテーションの対象となる談話単位の境界は文節の境界と一致すること，書き言葉を対象に文節のチャンキングを行った場合，例外的な現象を除いてほとんど正しく解析できるためである．このアノテーションの方法を採用することで，自動的に談話単位の境界を検出する場合にも，文節を単位とした処理を想定することになる．まず，以降で人手でどのように談話単位を検出するかについての基準について議論する．

1 節で示したように，談話関係解析のために談話単位を決定するためには，単純に節に相当する単位を談話単位とすれば良いわけではなく，談話単位間の関係を判断できるような命題相当の単位を検出する必要がある．このため，以下の基準 1.〜基準 5. の 5 つの談話単位認定基準を策定した．

基準 1. 文末は談話単位の境界とする．

基準 2. 補足節や連体修飾節の中にに含まれない従属節（並列節・副詞節）はその末尾を談話単位の境界とする．

基準 3. 補足節や連体修飾節の末尾は談話単位の境界としない．

基準 4. 機能語相当・副詞相当の表現の末尾は談話単位の境界としない．

基準 5. 文末に広義のモダリティに相当する表現が出現する場合，その表現を除外した後に基準 1. 〜 基準 4. の基準と照合して談話単位の境界を検出する．

まず，基準 1 の文末を談話単位の境界とする基準に関しては，命題を表す基本的な単位を文とみなすと自明である．次に，基準 3 の補足節や連体修飾節に関しては，例えば，(3) a. では，「おじいさんが山へ行った」「おばあさんが川へ行った」という 2 つの命題を想定すれば良いのに対し，(3) b. ではもし「おじいさんは山に行き」の後を境界とすると，「おじいさんは山へ行った」と「おばあさんは川に行ったと母が教えてくれた」という 2 つの命題に分割できるが，この場合，前者も後者の談話単位と同様に母が発話して伝達された内容という情報が捨象されることになる．この場合，この 2 つの談話単位間の談話関係を決定することが困難になるため，このように補足節に埋めこまれた節は分割しないこととした．

(3) a. [おじいさんは山に行き、][おばあさんは川に行った。][*2]

[*2] 以降では，1 つの談話単位の範囲を '[' と ']' で囲むことにより表

b. [おじいさんは山に行き、おばあさんは川に行ったと母が教えてくれた。]

連体修飾節の場合も同じような考え方に基づき，連体修飾節内は談話単位の分割対象から除外する．

次に，基準 2 であるが，これは基準 3 に該当しない節は独立に命題を構成できると考え，それらの間を談話境界として分割する．例えば，例 (3) a. の 2 つの節は埋め込み構造の中に出現していないため，「おじいさんは山へ行った」と「おばあさんは川へ行った」の 2 つを独立した談話単位とみなす．

また，「〜に関して」「〜によって」のように表現としては機能語・副詞相当だが，「関する」「よる」が品詞としては動詞に相当するため，品詞などの表層的な情報を用いて自動的に談話単位に分割すると分割してはならない箇所で分割される場合がある．例えば，例 (4) に出現する「に対して」は機能語相当表現であるが，場合によっては「対する」という動詞として解析されることがある．同様に，例 (5) の「違って」は副詞であるが，これも形容詞「違う」と品詞が付与されることで，談話単位の末尾となる可能性がある．この判断は言語学に関する知識を持つアノテーション作業者が作業する場合，判断を誤ることは少ないが，そうではない場合に解釈を誤る可能性が考えられる．このため，この判断に関して 4 つ目の基準を用意し，機能語・副詞相当なのか，それ以外なのかの判断を明示的に行うことにする．

(4) [新しい法律に対して見解を述べる。]

(5) [世界が違って見える。]

最後に，広義のモダリティに相当するために基準 5 を用意した．日本語では「〜する可能性がある」「と思われる」「とされる」などのモダリティ相当の言い回しでは，連体修飾節や補足節をともなうために，過度に節が埋め込まれることが頻繁に起こる．この結果，その埋め込まれた節の中は基準 3 のために談話単位の分割ができなくなる．この問題を回避するため，このような表現が出現した場合には，これらが出現していない状態で埋め込まれた談話単位を分割し，かつこの基準 5 に従って作業したことを表すために，このモダリティ相当の表現に「広義モダリティ」のタグを付与する．例えば，例 (6) や例 (7) では，それぞれ「と思う」「可能性がある」に「広義モダリティ」のタグを付与し，埋め込まれた節を必要に応じて談話単位に分割する．

(6) [毎年恒例であるから、][そういうものだと思う。]

(7) [株価が悪化し、][経済に打撃を与える可能性がある。]

これら 5 つの基準にしたがい，BCCWJ の一部 156 記事に関して談話単位の境界をアノテーションした．この結果，全体で 2,893 箇所に談話単位の境界情報が付与された．ジャンルごとの詳細は表 1 にまとめる．

現する．

表 1 BCCWJ の一部を対象としたアノテーションの結果

ジャンル	記事	文	文節	談話単位
PB:書籍	3	211	2,439	148
PM:雑誌	5	1,334	7,421	502
PN:新聞	11	542	3,855	186
LB:書籍	3	668	5,724	302
OB:ベストセラー	4	483	3,886	246
OW:白書	7	871	8,825	562
OL:法律	7	826	10,934	374
OC:Yahoo!知恵袋	102	669	3,756	418
OY:Yahoo!ブログ	14	209	1,650	155
合計	156	5,813	48,490	2,893

4. 人手による談話単位境界の特徴分析

3 節で作成した談話単位の境界がアノテーションされたコーパスを人手で分析することで，自動分割に利用するための特徴を考える．ここでは特に分割対象となる文節の内容語や機能語，またその文節から文末への係り受けの構造がどのように談話単位分割に関係するかについて分析する．

4.1 文節内の内容語（述語）に関する特徴

まず，談話単位認定基準 2 からわかるように，談話単位の境界となる文節は並列節や副詞節などの節末に相当するため，当該の文節には述語に相当する表現が含まれる．このため，例 (8)[*3]の「確立する」や例 (9) の「（私の）母だ」のような述語が存在することを捉えることで，談話境界の末尾であることがわかる．この特徴は文節の主辞の品詞や「⟨ 名詞 ⟩+だ」のような出現パタンを考慮することで判断可能である．

(8) [$_0$ 信念を $_1$ / $_1$ 確立し、$_3$] / [$_2$ 相手と $_3$/$_3$ 対決する。]

(9) [$_0$ そこに $_1$/$_1$ いるのが $_2$/$_2$ 私の $_2$/$_3$ 母で、$_6$] / [$_4$ 今年 $_5$/$_5$50 歳に $_6$/$_6$ なります。]

4.2 文節内の機能語に関する特徴

4.1 で示した内容語の情報に加え，文節内の機能語も談話単位の末尾を判断するための有益な手がかりを与える．例えば，例 (10) で動詞「遊ぶ」の後に接続助詞「で」が出現していることが，談話単位の末尾か否かを判断する手がかりとなる．

(10) [$_0$ 繰り返し遊んで、$_1$] / [$_1$ クリアできた。]

同様に，例 (11) では「わたる」が動詞ではなく，「にわた

[*3] 以降の例文では “/” を文節の区切り記号とし，また文節頭の下付き数字を文節の ID，文節末の下付き数字を係り先の ID として表記する．例えば，文節 ID が 1 の「確立し、」は文節 ID が 3 の「対決する。」に係ることを表す．

る」という連語に相当することが形態素解析の結果よりわかるので，その文節が末尾に該当しないことを表す手がかりとなる．

(11) [頭部を / 数回にわたり / 殴打した。]

さらに，例 (12) の「遊んで、」の文節のように，読点の出現が境界を表す手がかりとなる．この場合，係り受けの構造を考えることでも境界の認定は可能であるが，必ずしも係り受け解析が正しく解析できる保証はないため，このような記号に相当する手がかりも併用することでより正しい境界の推定を行うことができると考えられる．

(12) [$_0$ 繰り返し遊んで、$_3$] / [$_1$ クリアできる $_2/_2$ ゲームが $_3/_3$ 良いと $_4/_4$ 思った。]

4.3 係り受け関係に関する特徴

談話単位認定基準の基準 3 に定義したように，引用や連体修飾節の中は基本的には複数の談話単位に分割しない．例えば，例 (13) の「あるから」の後で分割しない理由は，この文節から文の末尾の文節の係り受けのパス中に引用の格助詞「と」が出現しているため，「予定があるから難しい」という表現が埋め込まれていることがわかるからである．

(13) [$_0$ 太郎は、$_4$ / $_1$ 予定が $_2$ / $_2$ あるから $_3/_3$ 難しい、と$_4$ / $_4$ 言った。]

同様に，例 (14) で「繰り返し遊んで」の後で分割し理由は，この文節から文の末尾までの係り受けのパス中に名詞「ゲーム」が出現していることから，当該の文節が連体修飾節の中に存在することがわかるからである．

(14) [$_0$ 繰り返し遊んで $_1/_1$ クリアできる $_2/_2$ ゲームが $_3/_3$ 良いと $_4/_4$ 思った。]

また，基準 5 で定義したように広義のモダリティの情報も係り受け関係から推定可能である．例えば，例 (15) の文の末尾には「と思う」という表現が文節を横断して出現しているが，この表現を広義のモダリティと考える場合，分類対象となる文節から文の末尾までにこの表現が出現していることが分割するか否かを判断する手がかりとなる．ここではこの「と思う」を広義のモダリティ相当とみなすため，「恒例であるから」の文節の末尾が談話単位の末尾となる．

(15) [$_0$ 毎年 $_1/_1$ 恒例であるから $_3$] / [$_2$ そういう $_3/_3$ ものだと$_4/_4$ 思う。]

さらに，例 (16) では形容詞「おいしい」が「いただく」に係っているが，この際に品詞が形容詞であるにもかかわらず「おいしく」という表現が副詞相当の役割を果たすため，この後で談話単位は分割されないことになる．このように，隣接する述語間の係り受け関係も分割のための手がかりとなる．

(16) [$_0$ 出された $_1/_1$ 料理を $_3/_2$ おいしく $_3/_3$ いただく]

表 2 談話単位分割に利用する素性

種類	素性
内容語	文節内の主辞の品詞・見出し語
	文節内の主辞の直後の形態素が助動詞「だ」か否か
機能語	文節内の機能語の品詞・見出し語・活用形
	文節末の読点の有無
係り受け	文節の主辞の品詞が名詞-非自立-副詞可能であり，その文節に係る文節の中に用言相当の語があるか否か
	文末の文節までの係り受けのパス中に主辞が名詞となる文節があるか否か
	文末の文節までの係り受けのパス中に品詞が助詞-格助詞-引用である語があるか否か
	文末の文節までの係り受けのパス中に品詞が助詞-格助詞-引用である語を含む文節があり，かつその直後の文節が動詞「思う」を含むか否か
	文末の文節までの係り受けのパス中に名詞と用言の両方を含む文節がある場合，その文節に係る文節の活用形が連体形か否か
	文節の係り先が直後の文節であるか否か

5. 提案する談話単位分割モデル

3 節で作成した談話単位がアノテーションされたコーパスには文節が談話単位の末尾であるか否かという情報がアノテーションされている．そこで，本研究では，談話単位の範囲を同定するのではなく，談話単位の末尾となる箇所を同定することで各談話単位の範囲を決定する．このため，解くべき問題は任意の文節が談話単位となるか否かを当てる 2 値分類問題となる．そこで，4 節に示した 3 種類の特徴を機械学習に基づく分類手法で利用するために，各特徴に関連する素性を抽出する．例えば，内容語や機能語に関する素性は分類対象となる文節の主辞や機能語の品詞や見出し語，活用形などを利用する．一方，係り受けに関する情報はいくつかの特徴的なパタンを捉える必要があるため，あるパタンに該当するか否かを素性として利用する．これらの素性の詳細を詳細を表 2 にまとめる．

6. 評価実験

6.1 実験設定

評価実験では 3 節で作成した談話単位境界タグ付きコーパスを用いて，5 分割交差検定による評価を行う．提案手法の学習・分類には SVM[10] を利用した．

提案手法の比較対象として 2 種類のベースラインモデルを採用した．一つは「機能語が接続助詞もしくは連用形の語」かつ「文節末に読点がある」文節を談話単位の末尾とする規則ベースの手法である．もう一つは，表 3 に示す丸山ら [5] が提案した節境界表現を含み，かつ述語に相当す

表 3 丸山ら [5] の提唱する節境界表現

強境界	ガ，ケド，ケドモ，ケレド，ケレドモ，シ
弱境界	タリ，テカラ，テハ，テモ，テ，トイウ，トカ，ノニ，ヨウニ，トノ，タラ，タラバ，ト，ナラ，ナラバ，レバ，ダノ，デ，ナリ，カラ，カラニハ，ノデ，テノ，連用形，引用助詞，感動詞

表 4 実験結果: クローズドテスト

モデル	精度	再現率	F 値
規則ベースのベースライン手法	0.817	0.504	0.624
丸山ら [5] の節境界分割手法	0.522	0.664	0.585
提案手法	0.814	0.811	**0.812**

表 5 ジャンル毎の評価結果

ジャンル	精度	再現率	F 値
PB: 書籍	0.829	0.793	0.811
PM: 雑誌	0.822	0.818	0.820
PN: 新聞	0.775	0.757	0.766
LB: 書籍	0.775	0.866	0.818
OB: ベストセラー	0.769	0.827	0.797
OW: 白書	0.885	0.839	0.861
OL: 法律	0.820	0.798	0.809
OC: Yahoo!知恵袋	0.913	0.757	0.828
OY: Yahoo!ブログ	0.899	0.799	0.846
全体	0.814	0.811	0.812

る文節を談話単位の末尾とする手法である．

6.2 実験結果

評価実験を行った結果を表 4 にまとめる．この結果より，ベースラインの 2 つの手法が F 値で約 0.6 という結果を得ているのに対し，提案手法ではそれを上回る F 値で 0.812 という結果を得た．これはベースラインとした手法が分類対象となる文節の局所的な情報しか参照していないに対し，提案手法ではそれらの情報に加え，係り受けのパスから得られる情報を追加したことで分類の性能が向上していると考えられる．

また，BCCWJ で規定されているジャンルごとの精度についても調査を行った．結果を表 5 にまとめる．この結果より，均質な言い回しで記述される白書や比較的平易な表現で記述されるブログが特に精度が良かったのに対し，複雑な入れ子構造を文内に持つ新聞記事などではその特徴を本研究で導入した素性では捉えることができなかったため，精度が低下したと考えられる．このことから，文の構造情報をジャンル横断的に扱えるようにするためにどのようにそれを抽象化し，どのように学習に利用するかを検討することが今後さまざまなジャンルに対して頑健なモデルを構築するために重要となると考えられる．

また，この実験で利用した評価用データは表 2 に示した素性を設計する際にも参照しているため，この評価実験はクローズドな評価でしかない．このため，未知のデータに対し提案手法がどの程度有効であるかを調査するために，クローズドテストで利用していない事例を BCCWJ から選択し，それを用いて評価実験を行う必要がある．そこで，BCCWJ から抜粋した事例集合に対し，新規に談話単位の境界をアノテーションを行った．このアノテーション作業の結果を表 6 にまとめる．

表 6 オープンテスト用のデータセット

ジャンル	記事	文	文節	談話単位
PB:書籍	2	210	2,227	146
PM:雑誌	1	200	1,062	42
PN:新聞	1	204	1,076	50
LB:書籍	2	200	1,588	116
OB:ベストセラー	2	205	1,992	131
OW:白書	7	203	2,007	113
OL:法律	2	208	2,953	89
OC:Yahoo!知恵袋	34	211	1,098	132
OY:Yahoo!ブログ	17	202	1,142	63
合計	68	1,843	15,145	882

表 7 実験結果: オープンテスト

モデル	精度	再現率	F 値
規則ベースのベースライン手法	0.781	0.492	0.604
丸山ら [5] の節境界分割手法	0.522	0.664	0.585
提案手法	0.779	0.789	**0.784**

この新規に作成した評価用データを利用し，オープンテストとなる評価実験を行った．結果を表 7 にまとめる．表 7 より，それぞれの結果の傾向は表 4 にまとめたクローズドテストの結果とほぼ同じであることがわかる．つまり，2 つのベースラインと比較して提案手法のほうが精度よく談話単位の境界を検出できている．ただし，クローズドテストの結果と比較して F 値で約 0.03 精度が低下しており，素性設計の際に出現していなかった事例が出現しているため結果が悪くなっていると考えられる． 6.3 節ではこの誤りの傾向を分析するために，解析を誤った事例のうち 150 事例をランダムに抽出し，その誤りの原因を分析する．

6.3 誤り分析

機械学習に基づく談話単位分割の結果解析を誤った事例から無作為に選択した 150 件を対象に人手で誤りの種類を調査した．この結果を表 8 にまとめる．以降で，それぞれの誤りについて説明する．

係り受け解析の誤り: 連体節や補足節に含まれるか否かを判別する際に，係り受け解析の結果に基づいてそれらを決定するが，この際，係り受け解析の結果が誤っているために，最終的な談話単位分割の結果に誤りが生じる．この誤りが最も多く，誤り全体の約 17%に相当する．

例えば，例 (17) では文節「導き出されるばかりではなく」は文節「あるのではないか、という」に係り，連体修飾節

表 8 解析誤りの分類

原因	個数	割合 [%]
係り受け解析の誤り	26	17
談話境界末尾を捉える語彙的情報の不足	21	14
括弧の挿入による文節境界判定の誤り	14	9
形態素解析の誤り	13	9
節末を捉える素性の不足	13	9
副詞相当表現の検出誤り	11	7
その他	42	28
分類不能	10	7
合計	150	100

の中に出現していることがわかるため，この文節の後で分割してはならない．しかし，係り受け解析の結果，この文節は文節「大事にしたい」に係り，連体修飾の関係とはならないため，結果的に分割され，誤った談話単位の境界を認定してしまう．

(17) [$_0$ また、$_{20/1}$ 正しい $_{2/2}$ 告白に $_{3/3}$ 達して、$_{7/4}$ そこから $_{7/5}$ 正しい $_{6/6}$ リアリティが $_{7/7}$ 導き出されるばかりではなく、$_{20/8}$ リアリティ $_{9/9}$ ある $_{10/10}$ フィクションに $_{11/11}$ 達して $_{15/12}$ はじめて $_{15/13}$ フィクションが $_{15/14}$ 完成品として $_{15/15}$ 認知される $_{16/16}$ 過程もまた $_{17/17}$ あるのではないか、という $_{18/18}$ 邪推を $_{20/19}$ いつも $_{20/20}$ 大事にしたい。]

このような問題を解決するためには，頑健な係り受け解析のシステムが必要となる．これまでに構築された係り受け解析器は新聞記事のみを学習事例として利用している場合がほとんどであり，その文章のスタイルを反映した学習を行っている可能性がある．頑健な係り受け解析のためには多様な現象が含まれた係り受けタグ付きコーパスが必要であるが，現状でそのようなものは存在しないため，そのコーパスを大規模に整備する，もしくは係り受け解析に依存しない談話境界の分割方法を考えるといった今後の方向性を検討する必要がある．

談話境界末尾を捉える語彙的情報の不足: 談話境界の末尾に出現する表現については，形態素の uni-gram だけではその多様性を捉えることができない．例えば，例 (18) のや「入れるだけで」や例 (19) の「やるのなら」は談話境界の末尾として検出することができなかった．このような「〈動詞〉するだけで」や「〈動詞〉するのなら」はこの形態素の並びが重要であるため，事前に表現を網羅し，その表現のパタンと合致するか否かを素性として導入することで対処したい．

(18) [$_0$ 俺は $_{11/1}$ 電柱の $_{2/2}$ 陰に $_{3/3}$ 隠れて $_{4/4}$ 見てるとさ、$_{11}$] / [$_5$ クマが $_{8/6}$ スポッと $_{8/7}$ 一発 $_{8/8}$ 入れるだけで $_{11}$] / [$_9$ 相手は $_{10/10}$ ドーッと $_{11/11}$ 倒れるだろ。]

(19) [$_0$ ＴＶで $_{2/1}$ 一緒に $_{2/2}$ やるのなら $_3$] / [$_3$ ゲームキューブですね。]

括弧の挿入による文節境界判定の誤り: 例 (20) の「(昭和十四年)」のように，書き言葉ではしばしば括弧による語や文の挿入が行われる．このため，この括弧の挿入がどのような情報の挿入であるかを捉えた上で，談話境界の分割を行う必要がある．この例では，「〜連載されたが」の補足的な情報として「(昭和十四年)」が記述されているため，この文節の末尾を談話境界の末尾として解析する必要がある．

(20) [$_0$ これは $_{12/1}$ 後に $_{3/2}$『救世軍士官雑誌』に $_{3/3}$ 連載されたが $_{12/4}$ (昭和 $_{5/5}$ 十四年)、$_9$] / [$_6$ 特に $_{9/7}$「軍人家族救護会」$_{9/8}$「捕虜の救護」$_{9/9}$ などは $_{12/10}$ 注目を $_{11/11}$ ひいた $_{12/12}$ ものである。]

形態素解析の誤り: 例 (21) の「で」は正しくは助動詞「だ」の連用形として解析されるべきだが，形態素解析器がこれを誤って格助詞「で」と解析してしまっている．このため，規則の適用対象から除外されることになり，結果的に誤った談話単位境界の判定を行ってしまう．

(21) [$_0$ 定員 $_{1/1}$ 二十人以上、$_{7/2}$ 三歳児未満の $_{3/3}$ 施設が $_{4/4}$ 対象で、$_7$] / [$_5$ 今では $_{7/6}$ 百八カ所に $_{7/7}$ 増えた。]

頑健に談話単位境界を認定するためには，このような形態素解析の誤りを許容できる仕組みを考えるべきであり，それも今後の課題となる．

副詞相当表現の検出誤り: 形態素解析の出力の性質上，副詞に相当する表現が他の用言に相当する品詞として解析される場合があり，その結果，その箇所が誤って談話境界の末尾として解析されることが起こる．

例えば，例 (22) で副詞相当の「連続して」が「連続する+て」と解析されるために，この文節が談話境界の末尾と認定されてしまう．

(22) [$_0$ たしか $_{1/1}$ 変な $_{2/2}$ 音の $_{3/3}$ しない $_{4/4}$ ボタンの $_{5/5}$ 場所を $_{6/6}$ 記憶して $_7$] / [$_7$ 連続して $_{8/8}$ 押さなくては $_{9/9}$ いけません。]

このような副詞相当の表現であるが，形態素解析の辞書に副詞として登録されていない語については，あらかじめ表現を列挙し，談話単位の末尾から除外することで対応したい．

述語認定の誤り: 例 (23) の文節「獲得、」のように，それ自体では名詞句なのか動詞句なのかが判別できない場合が存在する．

(23) [$_0$ この $_{1/1}$ 時は $_{4/2}$ 中曽根康弘さんが $_{4/3}$ 過半数を $_{4/4}$ 獲得、$_8$] / [$_5$ そのまま $_{8/6}$ 総理の $_{7/7}$ 座を $_{8/8}$ 射止めました。]

サ変名詞が名詞句なのか動詞句でかつ「する」が省略されているのかを判断するためには，係り先の文節が述語であるか否かが手がかりになると考えられる．この例の場合，

「射止めました。」という動詞句に係るため，係り元の文節「獲得、」も述語に相当する．本研究では係り先の情報を素性として利用したが，このような特徴までは考慮しなかったため，このような問題が起こった．今後は，上述のような特徴を素性に加えることで正しく分割できるかどうかを調査したい．

7. おわりに

本研究では，談話関係解析の前処理として必要となる談話単位分割の問題を解くために，さまざまな言語的な特徴に基づく分割手法を提案した．BCCWJ の一部を対象に人手で談話単位の境界をアノテーションし，これを学習・評価用データとして利用した．評価実験の結果，提案手法は交差検定によるクローズドテストで F 値が 0.812，オープンテストでは F 値が 0.784 という結果を得た．

今後の課題として，係り受け解析に過剰に依存しない談話境界認定の問題が考えられる． 6.3 節で示したように，係り受け解析の誤りが最も解析の誤りに影響していることがわかった．係り受け解析で典型的に間違う箇所はあらかじめ想定できると考えられるため，その箇所は解析に利用しない，もしくは利用するとしても解析の信頼度込みで利用するといった工夫が考えられる．

また， 3 節で導入した談話境界認定基準 5 の広義のモダリティ表現の扱いに関しては，「〜思う」という表現のみしか考慮していないため，今後はこの表現の収集に取り組む必要がある．まずは，日本語を対象とした言語学の分野で研究されているモダリティ表現を網羅的に収集し，それらを適用することで対応したい．

さらに，本研究では談話関係解析の前処理として談話単位分割の問題を扱ったが，その精度が十分に向上したのちに，この結果を前処理とした談話関係の解析に取り組む予定である．談話関係の定義には修辞構造理論 [11] など，さまざまな関係が提案されているが，それらの関係を吟味し，汎用的に使える関係のセットの模索，それらのアノテーションを行った上で，どの程度自動的に解析できるかを調査する予定である．

参考文献

[1] 武石 英二, 林 良彦. 接続構造解析に基づく日本語複文の分割. 情報処理学会論文誌, 33(5), pp. 652-663, 1992.
[2] 白井 諭, 瀬下 貴加子, 木村 淳子, 横尾 昭男, 池原 悟. 従属節の階層構造に基づく日本語長文の自動分割とその効果. 全国大会講演論文集, 第 53 回平成 8 年後期 (2), pp. 67-68, 1996.
[3] 白井 諭, 池原 悟, 横尾 昭男, 木村 淳子. 階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度. 情報処理学会論文誌, 36(10), pp. 2353-2361, 1995.
[4] 南 不二男. 現代日本語の構造. 大修館書店, 1974.
[5] 丸山 岳彦, 柏岡 秀紀, 熊野 正, 田中 英輝. 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, 11(3), pp. 39-68, 2004.
[6] 日本語話し言葉コーパス. http://www.ninjal.ac.jp/csj/
[7] 文部科学省科学研究費補助金特定領域研究　日本語コーパス. http://www.tokuteicorpus.jp/
[8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[9] 工藤 拓, 松本 裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, 43(6), pp. 1834-1842, 2002.
[10] V. N. Vapnik. Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing Communications, and control, 1998.
[11] W. C. Mann and S. A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. Text,8(3), pp. 243-281, 1988.
[12] L. Carlson, D. Marcu and M. E. Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pp. 1–10, 2001.