E-005

日本語節境界検出プログラム Rainbow の作成と評価

Rainbow: A New Detector of Japanese Clause-Boundaries

加納 隼人 † 佐藤 理史 †
Hayato Kanou Satoshi Sato

1 はじめに

日本語の文の構成要素には,「述語を中心としたまとまり」と定義される「節」という単位が存在する [1]．一般に，長い文は複数の述語を持つので，文を複数の節に分割することができる．このような節分割は，たとえば大学入試の国語の評論読解問題に現れるような長い文を解析する際に有用であると考えられる．

文中の節境界を検出するプログラムに，丸山らの開発した CBAP [2] がある．CBAP は，形態素解析済みテキストを入力とし，形態素列に対する局所的なパターンによって，節境界の位置と種類を検出する．検出に用いるルールは，人手で記述されたものである．

しかし，CBAP には以下の 2 つの問題点がある．まず，CBAP は Perl の正規表現を用いた文字列置換で実装されているため，可読性，拡張性に問題がある．次に，CBAP は文節境界を認定せずに節境界のみを認定するため，節境界検出ルールは文節境界に対する条件も含むものとなっており，過度に複雑である．

これらの問題点を解決するために，本研究では新たな節境界検出プログラム Rainbow を作成した．形態素解析済みテキストを入力とし，人手で記述したルールによって節境界の位置と種類を検出するという点では CBAP と同様であるが，Rainbow では，まず文節境界を検出し，次に節境界を検出するという手順を踏む．このような 2 段階検出を採用することにより，検出ルールの簡潔さと可読性・拡張性の向上をもくろむ．

2 Rainbow の構成

Rainbow の構成を図 1 に示す．形態素解析器には，MeCab/IPAdic を使用する．

2.1 文節境界の検出

Rainbow では，形態素間の境界のうち，左右の形態素が特定の品詞パターンとなっているものを，文節境界として検出する．文節境界は，比較的少数の簡潔なルールによって検出可能である．文節境界検出ルールの概要を表 1 に示す．

以下に，文節境界の検出例を示す．文節境界は「|」で表す．

| 太郎が | 重い | 荷物を | 軽々と | 運んだので |
花子は | 驚いた |

2.2 節境界

日本語における節とは,「述語を中心としたまとまり」と定義される [1]．上記の例文の述語は「運ぶ」と「驚く」であるので，節境界を「||」で表すと以下のようになる．

|| 太郎が | 重い | 荷物を | 軽々と | 運んだので ||
花子は | 驚いた ||

† 名古屋大学大学院 工学研究科 電子情報システム専攻

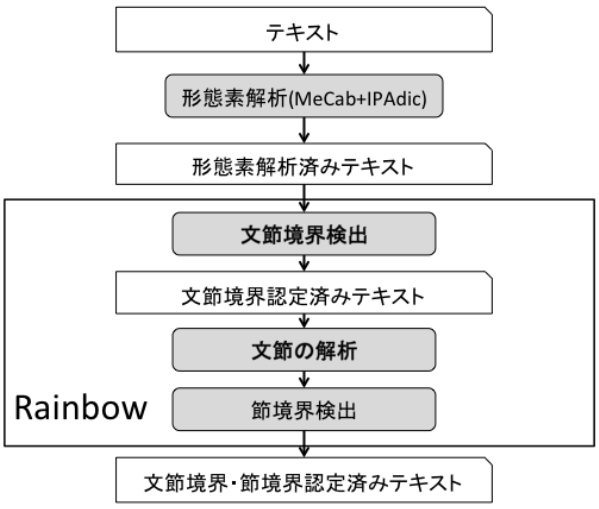

図 1: Rainbow の構成

表 1: 文節境界検出ルール

前の形態素	後ろの形態素	文節境界タイプ
助詞	(助詞 助動詞 以外)	助詞後
(名詞以外)	名詞	名詞前
(名詞以外)	動詞-自立	動詞前
副詞		副詞後
連体詞		連体詞後
	連体詞	連体詞前
接続詞		接続詞後
	形容詞-自立	形容詞前
助動詞	(助動詞 助詞 以外)	助動詞後
	読点	読点

このように，文節境界のうちのいくつかが節境界となる．なお，文頭と文末も便宜上，節境界とみなす．

日本語における節は主節と接続節の 2 つに分けられる [1]．原則として文末の述語を中心とした節を**主節**とよび，それ以外の節を**接続節**とよぶ．先の例文においては「花子が驚いた」が主節,「太郎が重い荷物を軽々と運んだので」が接続節である．さらに接続節は**従属節**と**並列節**に分けられ，従属節は**補足節，副詞節，連体節**の 3 つに分けられる．たとえば,「太郎が重い荷物を軽々と運んだので」は従属節であり，その下位分類は副詞節である．この節の大分類を表 2 に示す．

Rainbow では，主に接続節を対象として，節境界の検出を行う．同時に，益岡・田窪 [1] による節分類を参考にし，節の大分類の下にいくつかの小分類を設け，節の種類ラベルを付与する．以上の点は CBAP も同様であるが，節の小分類については CBAP と完全に同一ではなく，一部異なる．この小分類を表 3 に示す．括弧内の数字は，その分類に属するラベルの数を表している．これらに加え，便宜上<文頭>，<文末>，<主題> という節境

表 2: 節の大分類

節	主節		
	接続節	従属節	補足節
			副詞節
			連体節
		並列節	

表 3: Rainbow における節の小分類

大分類	小分類
補足節 (7)	形式名詞 (4)，疑問表現 (1)，引用 (2)
副詞節 (19)	時 (4)，原因・理由 (1)，条件・譲歩 (2)，付帯状況・様態 (2)，逆接 (1)，目的 (1)，程度 (1)，副詞節その他 (7)
連体節 (20)	連体節 (12)，内容節 (2)，時 (4)，疑問表現 (1)，形式名詞 (1)
並列節 (8)	逆接的 (1)，順接的 (7)

界を検出する．付与するラベルは全部で 57 種である．

2.3 文節の解析

節境界は，以下の 2 つのパターンに分けられる．

1. 述語を含む文節の直後が節境界となる場合
   例: || お腹が | **すいたので** || 饅頭を | 食べた ||
2. 述語を含む文節の直後に形式的な名詞を含む文節があり，その直後が節境界となる場合
   例: || 君が | あんなことを | **言った** | **せいで** || 大変な | ことに | なった ||

したがって，節境界の検出においては，文節が述語を含むかどうかが重要である．

Rainbow ではこの点に注目し，事前に文節の解析を行う．まずその文節が述語を含むか否かによって，述語文節，非述語文節に分け，その後，文節を述語の種類・活用形，含まれる名詞の種類，助詞の種類などの特徴の集合に変換する．すなわち，図 2 のような変換を行う．

2.4 節境界の検出

Rainbow では，文節境界の前後の文節列が特定の条件を満たす場合，その文節境界を節境界として認定する．

検出ルールの例を図 3 に示す．図 3 の上の例は，「基本形またはタ形の述語+接続助詞『が』」という形の述語文節を発見したら，<並列節/逆接的>というラベルを付与するルールである．下の例は，「基本形またはタ形の述語」という形の述語文節の後ろに，「非自立名詞『せい』+格助詞『で』」という形の非述語文節を発見したら，<副詞節/原因・理由>というラベルを付与するルールである．

節境界検出ルールの文節に対する条件は，文節解析によって得られた特徴集合に対して記述する．このことにより，節境界検出ルールの条件を，比較的抽象度の高いレベルで記述することが可能である．なお，ルール数は，130 程度である．

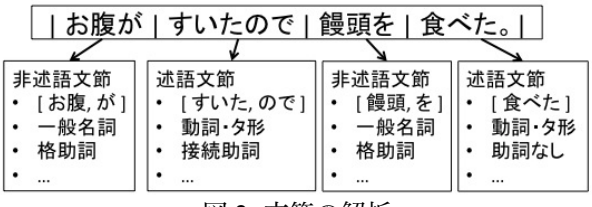

図 2: 文節の解析

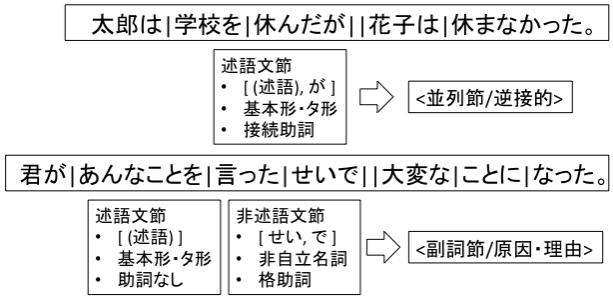

図 3: 節境界検出ルールの例

表 4: 評価結果

		位置のみ	位置と種類
Rainbow	*Precision*	89.3%	88.9%
	Recall	92.3%	91.8%
	F	**90.8%**	**90.3%**
CBAP	*Precision*	81.0%	77.3%
	Recall	97.1%	92.7%
	F	88.3%	84.3%

3 評価実験

人手で用意した正解データを用いて，Rainbow の評価を行った．正解データは，大学入試センター試験対策問題集「代々木ゼミナール 国語 2014」[3] の問題に現れる 140 個の選択肢に，節境界の位置と種類を人手で付与し作成した．節境界の位置のみ，および，位置と種類の両方を正しく検出しているかの 2 つの場合について，適合率 (*Precision*)，再現率 (*Recall*)，*F* 値を計算した．ただし，検出ルールが単純でほぼ確実に正解する<文頭>，<文末>，<主題> の節境界については，評価対象外とした．それらを除き，正解データには合計で 646 個の節境界が含まれる．

比較のため，同じ評価を Rainbow と CBAP に対して行った．Rainbow と CBAP では付与する節境界ラベルの小分類が異なるため，節境界の種類の検出においては，両者の小分類の共通する部分までの粒度において，正解，不正解を判定した．

評価結果を表 4 に示す．*Precision* は CBAP よりも Rainbow の方が高く，Rainbow の方が節境界の誤検出が少ない．しかし *Recall* は CBAP の方が高く，CBAP の方が節境界の検出漏れは少ない．*F* 値では，Rainbow は CBAP を上回った．

評価実験における Rainbow の検出誤りの原因は，(a) 節境界検出ルールの不備，(b) 形態素解析誤り，のいずれかであった．前者は，テストデータの解析結果をもとに節境界検出ルールを修正することで解決できる．後者は，たとえば「その中で生きようとして」の「として」を MeCab/IPAdic が格助詞と解析してしまうような場合である．このような誤りの一部は，文節解析により訂正できる可能性がある．

謝辞 本研究は，JSPS 科研費 24300052 の助成を受けた．

参考文献

[1] 益岡隆志, 田窪行則. 基礎日本語文法 -改訂版-. くろしお出版, 1992.

[2] 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 vol.11 No.3 pp.39-68, 言語処理学会, 2004.

[3] 代々木ゼミナール. 国語 2014 年版 (大学入試センター試験実戦問題集). 代々木ライブラリー, 2013.