BCCWJ に基づく中・長単位解析ツール

小澤俊介（電子化辞書班協力者：名古屋大学大学院情報科学研究科）†
内元清貴（電子化辞書班連携研究者：情報通信研究機構総合企画部）
伝康晴（電子化辞書班班長：千葉大学文学部）

Middle and Long Unit Word Analysis System Based on the BCCWJ

Shunsuke Kozawa (Graduate School of Information Science, Nagoya University)
Kiyotaka Uchimoto (Strategic Planning Department, NICT)
Yasuharu Den (Faculty of letters, Chiba University)

１．はじめに

文部科学省科学研究費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」の電子化辞書班では、語彙形態論研究に適した短単位、音声研究に適した中単位、構文・意味研究に適した長単位という複数粒度の「語」を高精度（98%以上）で自動構成するシステムを提供することを目的のひとつとしている。その各単位の例を図 1 と図 2 に挙げる。図 1 は「固有名詞仮名表記に関して論文を三本執筆した。」という文における短単位、中単位、長単位の関係を、図 2 は短単位と長単位の関係を表している。長単位は中単位を、中単位は短単位をそれぞれ結合することにより構成できる。例えば、「固有名詞仮名表記」という長単位は「固有名詞」「仮名表記」という二つの中単位から成るとともに、さらに「固有」「名詞」「仮名」「表記」のように分割した四つの短単位から成る。本稿では、中・長単位を自動構成する方法とそのツール Comainu について述べる。

文	固有名詞仮名表記に関して論文を三本執筆した。																							
文節	固有名詞仮名表記に関して								論文を			三本		執筆した。										
長単位	固有名詞仮名表記				に関して				論文		を	三本		執筆し			た	。						
中単位	固有名詞		仮名表記		に関して				論文		を	三本		執筆し			た	。						
短単位	固有	名詞	仮名	表記	に	関し	て		論文		を	三	本	執筆	し		た	。						

図 1：短単位、中単位、長単位の例

2. 長単位解析

2.1. チャンキングモデルと後処理に基づく長単位解析

長単位は、短単位列を入力とし、以下に述べるチャンキングモデルと後処理に基づく手法により認定する。長単位を認定するという問題は、長単位を構成する短単位のそれぞれに下記の四つのラベルのうちいずれかを付与する問題に置き換えることができる。これらのラベルの尤もらしさを推定するモデルを**チャンキングモデル**と呼ぶ。これは Uchimoto らの方法（Uchimoto & Isahara, 2007）におけるラベルの定義を次のように改良したものである。

† kozawa@el.itc.nagoya-u.ac.jp

短単位							ラベル	長単位					
書字形	語彙素読み	語彙素	発音形	品詞	活用型	活用形		書字形	語彙素読み	語彙素	品詞	活用型	活用形
固有	コユウ	固有	コユー	名詞-普通名詞-形状詞可能			**B**	固有名詞仮名表記	コユウメイシカナヒョウキ	固有名詞仮名表記	名詞-普通名詞-一般		
名詞	メイシ	名詞	メイシ	名詞-普通名詞-一般			**I**						
仮名	カナ	仮名	カナ	名詞-普通名詞-一般			**I**						
表記	ヒョウキ	表記	ヒョーキ	名詞-普通名詞-サ変可能			**I**						
に	ニ	に	ニ	助詞-格助詞			**B**	に関して	ニカンシテ	に関して	助詞-格助詞		
関し	カンスル	関する	カンシ	動詞-一般	サ行変格	連用形-一般	**I**						
て	テ	て	テ	助詞-接続助詞			**I**						
論文	ロンブン	論文	ロンブン	名詞-普通名詞-一般			**Ba**	論文	ロンブン	論文	名詞-普通名詞-一般		
を	ヲ	を	ヲ	助詞-格助詞			**Ba**	を	ヲ	を	助動-格助詞		
三	サン	三	サン	名詞-数詞			**B**	三本	サンホン	三本	名詞-数詞		
本	ホン	本	ホン	接尾辞-名詞的-助数詞			**I**						
執筆	シッピツ	執筆	シッピツ	名詞-普通名詞-サ変可能			**B**	執筆し	シッピツスル	執筆為る	動詞-一般	サ行変格	連用形-一般
し	シ	為る	スル	動詞-非自立可能	サ行変格	連用形-一般	**I**						
た	タ	た	タ	助動詞	助動詞-タ	終止形-一般	**Ba**	た	タ	た	助動詞	助動詞-タ	終止形-一般
。		。		補助記号-句点			**Ba**	。		。	補助記号-句点		

図 2：短単位と長単位の例

Ba　１短単位のみで長単位を構成し、かつ、その品詞、活用型、活用形が長単位のものと一致する。

Ia　長単位を構成する短単位のうち末尾の要素で、かつ、その品詞、活用型、活用形が

長単位のものと一致する。

B　長単位を構成する短単位のうち先頭の要素で、かつ、その品詞、活用型、活用形のいずれかが長単位のものと一致しない。

I　長単位を構成する短単位のうち先頭以外の要素で、かつ、その品詞、活用型、活用形のいずれかが長単位のものと一致しない。

これは、長単位を構成する末尾の短単位の品詞、活用型、活用形が長単位のものと一致する場合に付与されるラベルは「Ba」か「Ia」、そうでない場合は、長単位を構成する先頭の要素に付与されるラベルは「B」、長単位を構成する先頭以外の要素に付与されるラベルは「I」であることを意味する。したがって、このモデルにより、単位境界だけでなく、品詞、活用型、活用形の情報も得られる。例えば、図 2 の短単位には、「ラベル」の列に示されるようなラベルが付与される。これらのラベルを正しく推定できれば、「Ba」あるいは「Ia」が付与された短単位から品詞、活用型、活用形が得られる。図 2 は、「論文」などの長単位については品詞、活用型、活用形も得られることを表わしている。

チャンキングモデルの素性としては、着目する短単位とその前後 2 短単位、あわせて 5 短単位について、以下の情報を利用する。

- 書字形出現形、語彙素読み、語彙素表記、品詞、活用型、活用形、語種
- 階層化された素性（例えば「名詞-普通名詞-一般」）に対して、上位階層で汎化した素性（例えば「名詞」「名詞-普通名詞」）を利用する。
- 句読点や中点などの補助記号の場合、前後 1 短単位の品詞を「名詞-普通名詞-一般」「名詞-普通名詞-サ変可能」「名詞-普通名詞-副詞可能」「名詞-数詞」「その他」の 5 クラスにまとめ、素性として利用する。
- 辞書素性

  学習データ中の長単位から、2 短単位以上からなる助詞・助動詞について、前接する短単位の品詞・活用型・活用形と後接する短単位の品詞を含め（例えば、図 3)、辞書を作成した。辞書に含まれる長単位を構成する短単位列か否かの情報を素性として利用する。

この他、BCCWJ では①などの丸付き数字では長単位境界が区切れるため、丸付き数字か否かに関する素性も利用している。

一方、「執筆為る」などの長単位については品詞がこれらを構成する短単位「執筆」「為る」のどちらとも異なるため、各短単位には「B」あるいは「I」のラベルしか付与されない。この場合は、ラベルを正しく推定できたとしても品詞は得られず、単位境界の情報のみが得られることになるため、その長単位に対し、後処理として次に述べる品詞推定モデル及び活用型・活用形推定モデルを適用することにより、最も尤もらしい品詞、活用型、活用形を推定する。品詞推定モデルは、長単位を構成する短単位列が与えられると、学習データに現れた品詞を候補としてその品詞候補すべてについて尤もらしさを計算するモデルである。

			動詞	サ行変格	連用形
て	テ	て	助詞-接続助詞		
いる	イル	居る	動詞-非自立可能	上一段-ア行	終止形-一般
			助詞		

図 3：助動詞辞書の要素例

ただし、助詞と助動詞については長単位を構成する短単位列が複合辞と一致している場合のみ品詞候補とし、それ以外の場合には、助詞と助動詞を除くすべての品詞候補から最尤の品詞を出力する。複合辞と一致しているかどうかは、予め用意した複合辞辞書との文字列マッチングにより自動判定する。素性としては、着目している長単位とその前後の長単位、あわせて 3 長単位について、各長単位を構成する短単位の情報を用いる。具体的には、各長単位を構成する短単位について、先頭から 2 短単位と末尾から 2 短単位に着目し、各短単位に関する書字形出現形、語彙素読み、語彙素表記、品詞、活用型、活用形、及び、階層化された素性に対して上位階層で汎化した情報を素性として用いる。長単位が 1 短単位からなる場合は、先頭から 2 短単位目の情報は与えられなかったもの（NULL）として扱う。例えば、図 2 の「執筆し」では、「三本」「執筆し」「た」の 3 長単位に対し、「三|本|執筆|し|た|NULL」（先頭から各 2 短単位）、及び、「NULL|た|し|執筆|本|三」（末尾から各 2 短単位）の各短単位に関する情報を素性として用いる。活用型推定モデル、及び、活用形推定モデルは、推定するカテゴリが品詞ではなくそれぞれ活用型、活用形となる点、及び、動的素性を用いる点を除いて品詞推定モデルと同様である。動的素性としては、活用型推定モデルでは着目している長単位の品詞（自動解析時は品詞推定モデルにより自動推定した品詞）を、活用形推定モデルでは着目している長単位の品詞と活用型（自動解析時は品詞推定モデル、活用型推定モデルによりそれぞれ自動推定した品詞と活用型）を用いる。

長単位の語彙素読み・語彙素表記は、基本的に短単位の語形と語形代表表記をそれぞれ結合することで生成する。ただし、語彙素読みでは長単位末尾の短単位が活用語の場合は語形基本形を結合し、語彙素表記では長単位の末尾以外の短単位が活用語の場合は語形代表表記出現形を結合する。また、短単位の品詞が人名・地名になっている部分は短単位書字形を用いる。複合辞辞書に登録されている複合辞については、辞書引きにより読みと表記の情報を得る。

2.2. 実験と考察

2.1 節に述べた手法を用いて実験を行った。チャンキングモデルの学習と適用には、Yamcha と CRF++、MMA を用いた。Yamcha は SVM に基づく汎用チャンカーであり、カーネルは多項式カーネル（べき指数 3）を採用した。解析方向は文末側から文頭側とし、多クラスへの拡張は one-versus-rest 法を用いた。CRF++は CRF に基づく汎用チャンカーであり、MMA（Kruengkrai et al., 2009）は MIRA に基づく形態素解析システムである。後処理には SVM を用いた。BCCWJ の白書・書籍・新聞・雑誌・Web（Yahoo! 知恵袋）コアデータのうち、27,610 文（白書 : 5,216 文/205,150 短単位、書籍 : 8,288 文/212,878 短単位、新聞 : 14,101 文/326,402 短単位、雑誌 : 10,800 文/218,636 短単位、Web : 5,725 文/99,917 短単位）でモデ

ルを学習し、3,069 文（白書 : 580 文/23,127 短単位、書籍 : 921 文/21,656 短単位、新聞 : 1567 文/34,425 短単位、雑誌 : 1,200 文/26,911 短単位、Web : 637 文/10,835 短単位）で評価した。

表 1 に長単位解析の解析精度[1]を示す。白書・書籍・新聞・雑誌・Webのいずれに対しても、98%超の正解率となっている。また、CRF・MIRAのいずれのモデルを用いた場合でも、語彙素認定において 98%を超える精度が得られている。

表 1：長単位解析システムの解析精度

モデル		白書	書籍	新聞	雑誌	Web	全て
CRF	境界認定	99.3	99.0	98.9	98.7	98.4	98.9
	品詞認定	99.1	98.8	98.6	98.4	98.3	98.6
	語彙素認定	99.0	98.6	98.6	98.4	98.3	98.6
MIRA	境界認定	99.3	98.9	98.9	98.7	98.5	98.9
	品詞認定	99.0	98.7	98.5	98.4	98.4	98.6
	語彙素認定	99.0	98.6	98.5	98.3	98.4	98.5

3. 中単位解析

中単位は語の内部構造に従った単位であり、長単位を超えない範囲で、直接的な係り受け関係を持つ、隣接する短単位同士を結合したものとして定義できる。中単位は、長単位を入力とし、以下に述べる短単位間の係り受け解析と中単位境界同定ルールにより認定する。例えば、図 4 の 4 短単位から構成される長単位「固有名詞仮名表記」では、「固有名詞」と「仮名表記」の 2 つの中単位が生成される。

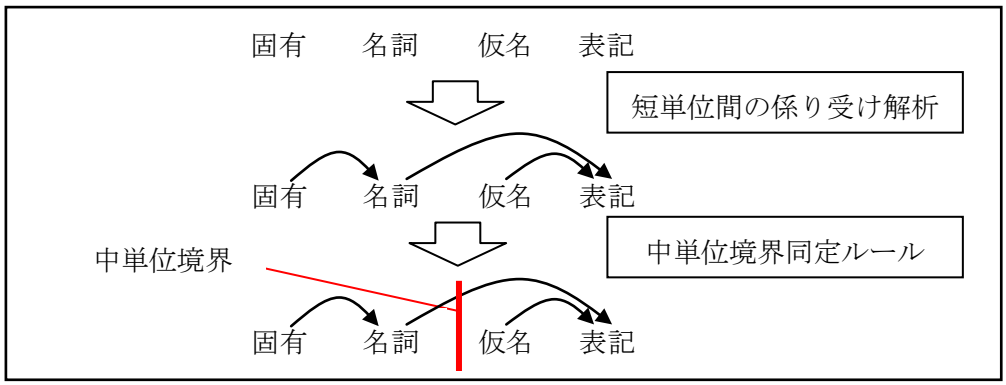

図 4：中単位解析の例

3.1. 短単位間の係り受け解析

最大全域木に基づく依存構造解析手法（McDonald et al., 2005）を短単位間の係り受け解析に適用した（Uchimoto & Den, 2008）。学習および解析には MST Parser を用いた。first order

[1] SVM の学習に時間がかかったため、予稿集には間に合いませんでした。最新の結果はポスター発表をご参照ください。

の素性を採用した。短単位間の係り受け解析の素性には、以下の情報を利用する。

- 書字形出現形、語彙素表記、品詞、活用型、活用形
- 階層化された素性に対して、上位階層で汎化した素性

BCCWJ の白書・書籍・新聞コアデータのうち、6,547 文（白書：2,068 文/81,867 短単位、書籍：3,308 文/82,175 短単位、新聞：1,171 文/29,161 短単位）に対して、短単位間の係り受け情報と中単位情報を付与したデータを対象に学習、評価した。学習および評価は 10 分割交差検定により行った。このときの係り受け解析精度を表 2 に示す。白書・新聞に対しては約 98%、書籍に対しては 99%超の正解率となっている。ただし、長単位を構成する短単位数が増加するにつれて精度が低下するため、長い長単位について今後改善が必要である。

表 2：短単位間の係り受け解析精度

長単位を構成する短単位数	白書	書籍	新聞	全て
2 短単位以上	97.9	99.4	97.7	98.3
3 短単位以上	97.9	97.7	94.9	96.3
4 短単位以上	94.0	96.2	92.4	94.0
5 短単位以上	92.3	95.7	91.5	92.5

3.2. 中単位境界同定ルールによる中単位の認定

短単位間の係り受け情報に基づく中単位境界同定ルールにより、中単位境界を認定する。以下にルールを示す。

1. 長単位を超えない範囲で、順次係り受けの語は繋げる。
2. 語をまたいだ係り受けは区切る。
3. 補助記号は前後の形態素と区切る。

ただし、例外として以下のルールを設ける。

- 長単位の品詞が形状詞の場合
  - 語をまたいだ係り受けの場合も区切らず、一つの中単位とする。
- 長単位の品詞が名詞の場合
  - 短単位が以下の接頭辞の場合は区切る。
    各、計、現、全、非、約
  - 係り受けが並列の場合、並列の形態素同士は区切る。但し、並列の形態素の品詞が接頭辞の場合は、区切らない。
  - 後続する短単位列が「名詞+接尾辞」であり、接尾辞に係る場合は区切らない。
  - 後続する短単位列が「接頭辞+名詞」であり、名詞に係る場合は区切らない。

3.2 節で述べた短単位間の係り受け情報を自動付与したデータを対象に、中単位認定を行った。表 3 に中単位解析の解析性能を示す。白書・書籍・新聞のいずれに対しても、F 値で 98％～99％となっている。しかし、白書と新聞では長単位を構成する短単位数の増加に伴

う性能低下が著しい。これは次に述べる理由から短単位間の係り受け解析による影響と考えられる。

表 3：中単位解析の解析性能（F 値）

長単位を構成する短単位数	白書	書籍	新聞	全て
全て	98.5	99.8	98.9	99.2
2 短単位以上	95.2	99.2	96.3	96.6
3 短単位以上	91.3	97.0	91.6	92.3
4 短単位以上	83.1	95.6	84.2	85.0
5 短単位以上	77.2	95.8	85.3	81.7

中単位境界同定ルールの性能を調べるため、正解の係り受け情報を用いて中単位境界解析を行った。その結果、長単位を構成する短単位数がいずれの場合であっても性能は 99%を超えており、中単位境界同定ルールは十分な性能を保持していることが分かった。このことから、中単位境界解析の性能を上げるには、短単位間係り受け解析の性能を上げる必要があると言える。

4. 中・長単位解析ツール Comainu

2 章と 3 章で説明した手法を実装することにより、中・長単位解析ツール Comainu を作成した。本ツールは以下の機能を持つ。

- 長単位解析
  平文または短単位列を入力すると、長単位を付与した短単位列を出力することができる。平文が入力された場合、Chasen もしくは Mecab により形態素解析を行った後に長単位解析を行う。長単位解析のチャンキングモデルには SVM と CRF、MIRA のいずれかを用いることができる。
- 中単位境界解析
  平文または短単位列もしくは長単位情報を付与された短単位列を入力すると、中・長単位を付与した短単位列を出力することができる。平文が入力された場合には形態素解析と長単位解析、短単位列が入力された場合には長単位解析を行った後に中単位境界解析を行う。
- 文節境界解析
  平文または短単位列を入力すると、文節境界を付与した短単位列を出力することができる。平文が入力された場合、形態素解析を行った後に文節境界解析を行う。

平文や短単位列の直接入力だけでなくファイル入力にも対応している。解析結果をファイルに保存することも可能である。

図 6 に Comainu による中・長単位解析の実行例を示す。図 6 の例では、短単位列を入力とし、MIRA を用いて学習したチャンキングモデルによる長単位解析及び中単位解析を実

行し、中・長単位が付与された短単位列を出力している。出力の 29～34 列目はそれぞれ長単位の書字形出現形、語彙素読み、語彙素表記、品詞、活用型、活用形を表し、出力の 35 列目は短単位間の係り受け情報、36 列目は中単位番号、37 列目は中単位の書字形出現形を表す。

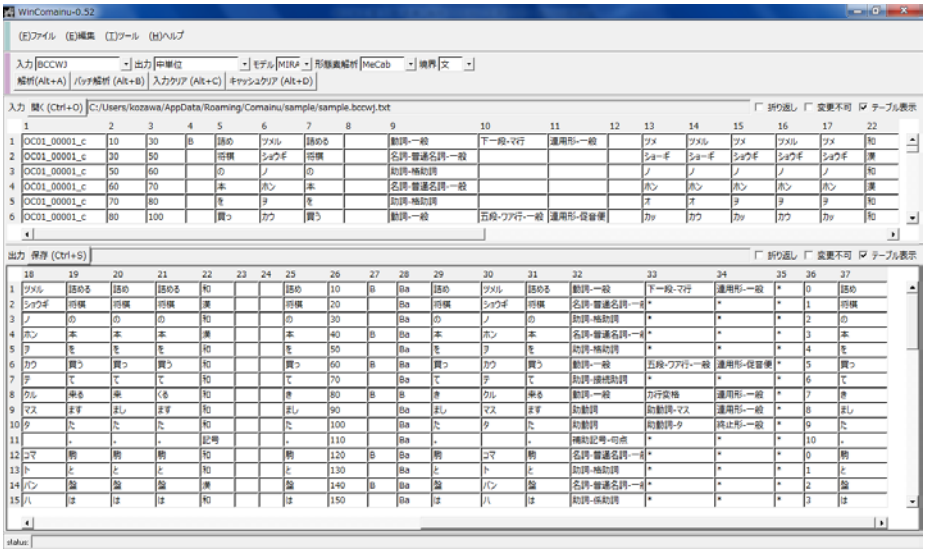

図 6： Comainu による中・長単位解析の実行例

5. まとめ

本稿では、チャンキングモデル及び後処理に基づく長単位解析手法、及び、短単位間係り受け解析及び中単位境界同定ルールに基づく中単位解析手法について述べ、それらを実装したシステムである中・長単位解析ツール Comainu を紹介した。BCCWJ コアを用いた実験では、境界認定 98.9%、品詞認定 98.6%、語彙素認定 98.6%の長単位解析精度、及び、99.2%の中単位解析精度を得た。今後、本ツールを公開する予定である。

文献

Kruengkrai, C., Uchimoto, K., Kazama, J., Wang, Y., Torisawa, K., & Isahara, H. (2009). An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging. In *Proc. of ACL-IJCNLP* 2009.

McDonald, R., Crammer, K., & Pereira, F. (2005). “Online Large-Margin Training of Dependency Parsers”, *In Proc. of ACL 2005*, pp. 91-98.

Uchimoto, K. & Den, Y. (2008). Word-level Dependency-structure Annotation to Corpus of Spontaneous Japanese and Its Application, In *Proc. of LREC* 2008.

Uchimoto, K., & Isahara, H. (2007). Morphological annotation of a large spontaneous speech corpus in Japanese. In *Proc. of the 20th IJCA I*, pp. 1731-1737.