BCCWJ に基づく長単位解析ツール

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）

Long Words Analysis System based on 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、この三種類の語の単位のうち、長単位を対象に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次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(Uchimoto & Isahara, 2007)。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。

Ba　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先頭以外の要素で、かつ、その品詞、活用型、活用

† kozawa@el.itc.nagoya-u.ac.jp

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。

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-サ変可能		
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			**Ia**						
に	ニ	に	ニ	助詞-格助詞			**Ba**	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			**Ba**	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

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、長単位を構成する先頭の要素に付与されるラベルは「Ba」か「B」であり、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「Ia」か「I」であることを意味する。また、「Ba」「Ia」が付与された要素は長単位と同じ品詞、活用型、活用形を持つことを意味する。したがって、このモデルにより、単位境界だけでなく、多くの場合、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、「執筆為る」以外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。一方、「執筆為る」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ため、下記に述べる後処理により品詞・活用型・活用形を推定する。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、発音形出現形、語彙素読み、語彙素表記、品詞、活用型、活用形
- 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。

				動詞	サ行変格	
て	テ	て	テ	助詞-接続助詞		
いる	イル	居る	イル	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

上記にあげた「執筆為る」の例のように、長単位を構成する短単位に付与されるラベルが「B」または「I」のみの場合、その長単位に対し、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、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、助詞と助動詞以外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。ここで、品詞候補は学習デ

ータに現れた品詞とする。ただし、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には，助詞と助動詞についても尤もらしさを計算する。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・発音形出現形・語彙素読み・語彙素表記・品詞・活用型・活用形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[1]は、基本的に短単位の語彙素読み・語彙素表記をそれぞれ結合することで生成する。ただし、活用語や複合辞については活用テーブルを参照することで読みと表記の情報を得る。一部の複合辞、つまり、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。また、品詞が「名詞-固有名詞-地名-一般」もしくは「名詞-固有名詞-地名-国」である短単位では規程により語彙素表記が片仮名表記となっているため，これらを構成要素に持つ長単位の場合、短単位の語彙素表記を単純に結合すると「カントウ地震」や「オウシュウ連合」のように片仮名表記を含む語彙素表記が生成されてしまう。そのため、地名を含む長単位については語彙素表記の代わりに書字形表記を利用する。

2.2. 実験と考察

Uchimoto らの方法（Uchimoto & Isahara, 2007）を改良したシステム（伝ほか, 2009）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・新聞コアデータのうち, 27,610 文（白書:5,217 文/205,520 短単位, 書籍:8,289 文/213,282 短単位，新聞：14,104 文/326,428 短単位）でモデルを学習し，3,069 文（白書：580 文/23,141 短単位，書籍：921 文/21,247 短単位，新聞：1568 文/34,445 短単位）で評価した。

---

[1]長単位の語彙素読み・語彙素表記については、語形レベルでの代表表記データが完成しており、語形出現形・語形出現形代表表記を単純に連接するだけで、長単位語彙素読み・語彙素表記を得られるようになっている。したがって、今後は語形の情報を用いる予定である。

表 1 に長単位解析の解析精度を示す。白書・書籍・新聞のいずれに対しても，97%～98%超の正解率となっている。また、SVM・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。

図 4 に学習データ量と長単位解析精度との関係を示す。モデルは SVM を用いて学習した。図から、350 文程度の学習データ量でも境界認定精度は 95%と高く、1,400 文程度の学習データ量になると語彙素認定で 96%を超えること，学習データ量が増えるとさらに精度が向上していることが分かる。

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	全て
SVM[2]	境界認定	99.7	99.5	99.2	99.4
	品詞認定	99.5	99.2	98.8	99.1
	語彙素認定	98.7	98.6	97.5	98.1
CRF	境界認定	99.3	99.4	98.6	99.1
	品詞認定	99.0	99.2	98.2	98.7
	語彙素認定	98.5	98.7	97.2	98.1
MIRA	境界認定	99.6	99.3	98.9	99.2
	品詞認定	99.2	98.6	98.3	98.7
	語彙素認定	98.9	98.5	97.3	98.1

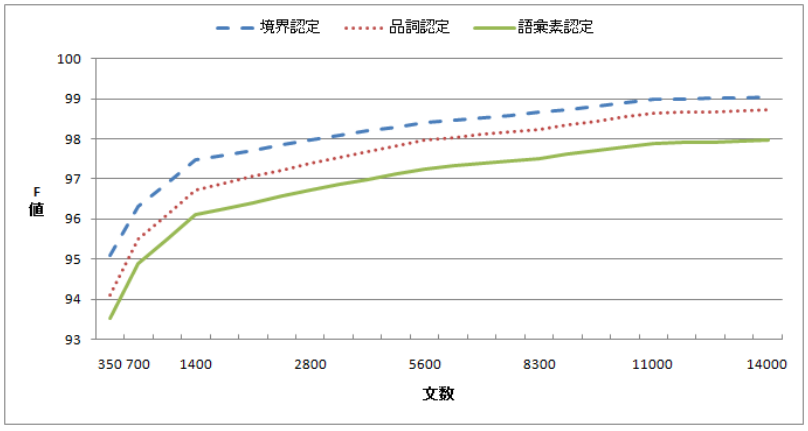

図 4：学習データ量と長単位解析精度の関係

3. 長単位解析ツール Comainu

2 章で説明した手法を実装することにより、長単位解析ツール Comainu を作成した。本ツールは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

[2] SVM の解析性能は、（冨士池ほか，2010）のものであり、最新の結果とは異なります。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。

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。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 5 に Comainu による長単位解析の実行例を示す。図 5 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り長単位解析を実行し、長単位が付与された短単位列を出力している。出力の 17～19 列目はそれぞれ長単位の品詞、活用型、活用形を表し、23～25 列目はそれぞれ長単位の語彙素読み、語彙素表記、書字形出現形を表す。

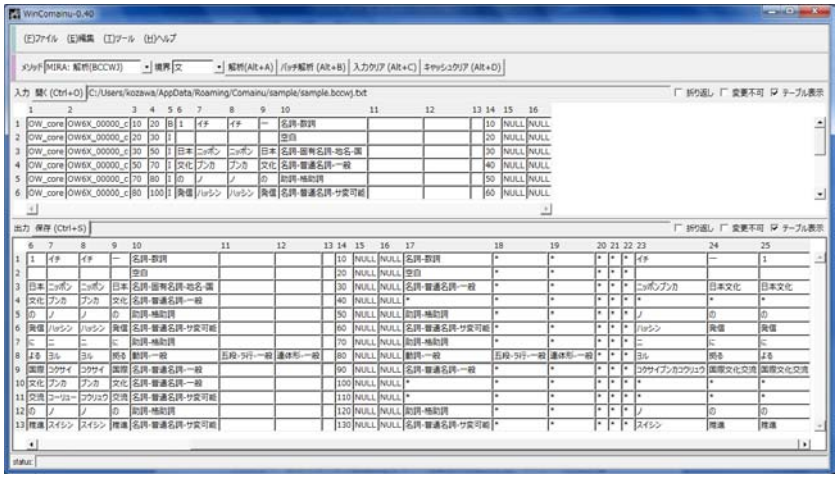

図 5： Comainu による長単位解析の実行例

4. 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複数の統計学習手法に基づく長単位解析ツール Comainu について述べた。BCCWJ コア（白書、書籍、新聞）を用いた実験では、境界認定 99.2%、品詞認定 98.7%、語彙素認定 98.1%の解析精度を得た。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.

Uchimoto, K., & Isahara, H. (2007). Morphological annotation of a large spontaneous speech corpus in Japanese. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence* pp. 1731-1737.

伝康晴・内元清貴・山田篤・峯松信明(2009).「UniDic 短単位解析以後の処理について」『特定領域研究「日本語コーパス」平成 21 年度全体会議予稿集』pp.85-92

冨士池優美・小椋秀樹・小西光・小木曽智信・小磯花絵・内元清貴・小澤俊介(2010).「『現代日本語書き言葉均衡コーパス』における長単位解析の進捗状況」『特定領域「日本語コーパス」平成 21 年度公開ワークショップ(研究成果報告会)予稿集』