BCCWJに基づく長単位解析ツール Comainu

小澤 俊介 † 内元 清貴 ‡ 伝 康晴 §
(株) はてな †, 情報通信研究機構 ‡, 千葉大学 §
skozawa@hatena.ne.jp, uchimoto@nict.go.jp, den@cogsci.l.chiba-u.ac.jp

1 はじめに

近年，言語研究において，言語現象を統計的に捉えるため，コーパスを用いた研究が盛んに行われている．コーパスを用いた研究は，語法，文法，文体に関する研究 [3]，語彙に関する研究 [5]，時代ごとの言語変化を調査する通時的な研究 [4] など多岐にわたる．

現代日本語書き言葉均衡コーパス（以下，BCCWJ）では言語単位として語彙形態論研究に適した短単位，及び，構文・意味研究に適した長単位を用いている．言語的特徴の解明に適した長単位はコーパスを用いた研究において重要な単位である．そのため，BCCWJとは異なるコーパスに対しても長単位情報を自動付与できることが求められる．

本論文では，長単位情報を自動付与する手法，及び，その手法を実装したツール Comainu について述べる．

2 長単位解析手法

長単位解析とは長単位境界，及び，長単位の語彙素，語彙素読み，品詞，活用型，活用形を同定するタスクである．短単位解析では，辞書を用いることで，高精度に解析が行われきた [1]．長単位解析でも長単位辞書を構築することによって高精度に解析できることが考えられるが，辞書の構築には膨大な労力が必要となるため，効率的でない．そのため，短単位情報を組み上げることにより長単位解析を行う．

短単位と長単位の例を表 1 に挙げる．表 1 は「固有名詞に関する論文を執筆した。」という文における長単位，短単位の関係を表している．例えば，「固有名詞」という長単位は「固有」と「名詞」の 2 短単位から構成される．

2.1 チャンキングモデル

短単位列に対して，チャンキングモデルによって長単位境界を認定する．長単位解析では長単位の境界を認定するだけでなく，長単位の品詞情報なども認定する必要があるため，Uchimoto ら [2] が定義したラベルに改良を加えた下記の 4 つのラベルを用いる．

表 1: 文，長単位，短単位の関係

文	固有名詞に関する論文を執筆した。
長単位	固有名詞／に関する／論文／を／執筆し／た／。
短単位	固有／名詞／に／関する／論文／を／執筆／し／た／。

Ba 単独で長単位を構成する短単位で，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 複数短単位で構成される長単位の末尾の短単位で，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 複数短単位で構成される長単位の先頭の短単位．もしくは，単独で長単位を構成する短単位で，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 複数短単位で構成される長単位の先頭でも末尾でもない短単位．もしくは，複数短単位で構成される長単位の末尾の短単位で，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．

単独の短単位から構成される長単位に対しては，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は「Ba」，一致しない場合には「B」が付与される．一方，複数短単位から構成される長単位に対しては，先頭の短単位には「B」，先頭でも末尾でもない短単位には「I」，末尾の短単位にはその品詞，活用型，活用形が長単位のものと一致する場合には「Ia」，一致しない場合には「I」が付与される．

図 1 に「固有名詞に関する論文を執筆した。」に対して，ラベルを付与した例を示す．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 1

短単位列						ラベル	長単位列					
書字形	語彙素読み	語彙素	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	名詞-普通名詞-形状詞可能			B	固有名詞	コユウメイシ	固有名詞	名詞-普通名詞-一般		
名詞	メイシ	名詞	名詞-普通名詞-一般			Ia						
に	ニ	に	助詞-格助詞			B	に関して	ニカンスル	に関する	助詞-格助詞		
関する	カンスル	関する	動詞-一般	サ行変格	連体形-一般	I						
論文	ロンブン	論文	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	助詞-格助詞			Ba	を	ヲ	を	助詞-格助詞		
執筆	シッピツ	執筆	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	スル	為る	動詞-非自立可能	サ行変格	連用形-一般	I						
た	タ	た	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 1: 短単位と長単位の例

は,「に関する」「執筆し」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,「に関する」「執筆し」についてはこれらを構成する末尾の短単位が長単位の品詞と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などは得られず，単位境界の情報のみが得られることになるため，次節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，語彙素読み，語彙素，品詞，活用型，活用形，語種情報
- 階層化された素性に対して上位階層で汎化した素性をチャンキングモデルの素性として追加する．例えば「名詞-普通名詞-一般」に対しては,「名詞」「名詞-普通名詞」を素性として追加する．

この他，BCCWJ では①などの丸付き数字で長単位境界が区切れるため，囲み情報に関する素性も利用する．

2.2 カテゴリ推定モデルによる後処理

チャンキングモデルによりラベルを正しく推定できた場合，図 1 の「に関する」及び「執筆し」の各短単位には「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などは得られず，単位境界の情報のみが得られることになる．これらの長単位に対してはカテゴリ推定モデルによって品詞，活用型，活用形を付与する．

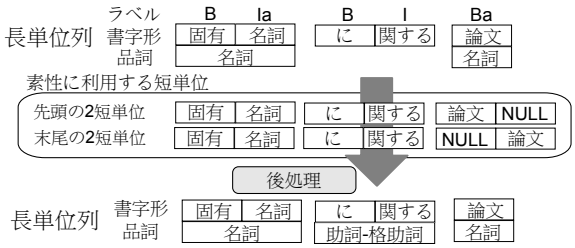

図 2: 品詞推定モデルの適用例

カテゴリ推定モデルは，学習データに現れたカテゴリを候補として，その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．長単位を構成する短単位列を与えると，その長単位に対して最尤のカテゴリを出力する．推定するカテゴリを品詞，活用型，活用形とした，品詞推定モデル，及び，活用型推定モデル，活用形推定モデルをそれぞれ学習・適用し，最も尤もらしい品詞，活用型，活用形を推定する．

推定するカテゴリを品詞とする品詞推定モデルでは，学習データに現れた品詞のうち，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．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候補とする．複合辞と一致しているかどうかは，複合辞辞書との文字列マッチングにより自動判定する．複合辞辞書は BCCWJ で認定された複合辞を予め人手で整理することにより用意した．

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 長単位に対して，先頭から 2 短単位と末尾から 2 短単位の計 12 短単位の情報を用いる．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．各短単位に対して，書字形，語彙素読み，語彙素，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．図 1 の「に関する」に対して品詞推定モデルを適用する例を図 2 に示す.「に関する」では，前後の長単位をあわせた「固有名詞」「に関する」「論文」の 3 長単位に対し，先頭と末尾の 2 短単位の情報を素性として用いる．図 2 では，最尤の品詞として助詞-格助詞を出力している．

活用型推定モデル，及び，活用形推定モデルは，推

表 2: BCCWJ を学習したモデルを用いた実験結果

	モデル	精度 (%)	再現率 (%)	F 値
境界推定	ベースライン	98.76	98.76	98.74
	提案手法	98.94	98.92	98.93
品詞推定	ベースライン	97.66	97.71	97.68
	提案手法	98.67	98.65	98.66

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．

3 実験と考察

3.1 実験

提案手法の性能を示すため，BCCWJ を用いた実験を行った．実験で用いるデータは BCCWJ のコアデータの一部を学習データとテストデータに分け，学習データには 18,140 文（332,009 長単位，419,414 短単位），テストデータには 2,015 文（36,297 長単位，45,906 短単位）を用いた．また，ベースラインとして Uchimoto ら [2] の手法を用いた．

チャンキングモデルの学習と適用には，CRF++[1]を用いた．パラメータは CRF++のデフォルトのパラメータを用いた．また，改良した後処理に用いる品詞，活用型，活用形推定モデルの学習には YamCha[2]を用いた．カーネルは多項式カーネル（べき指数 3）を採用し，多クラスへの拡張は one-versus-rest 法を用いた．

実験の結果を表 2 に示す．境界推定，及び，品詞推定のどちらにおいてもベースラインより高い性能を示し，境界推定では 98.93%，品詞推定では 98.66% の性能が得られた．

3.2 考察

3.2.1 誤り傾向の分析

まず，境界推定誤りの傾向を調査したところ，2 つの誤りの傾向が見られた．1 つ目は名詞連続であり，名詞の短単位列に対する長単位境界を誤ることが多かった．例えば，「医学部倫理委員会」では「医学部」と「倫理委員会」の 2 つの長単位にすべきところを「医学部倫理委員会」と誤って 1 長単位として判定した．これらを正しく解析するには，「部」や「庁」などの境界になりやすい短単位の辞書を構築し，境界になりやすい短単位かどうかを素性として追加することである程度対応できるだろう．2 つ目は複合辞に関する誤りであり，複合辞相当の長単位を複合辞として判定できなかったり，逆に複合辞ではない短単位列を複合辞として判定してしまうことが多かった．誤りの例を表 3 に示す．表の例では，長単位境界を「|」で表している．短単位列が複合辞か否かは前後の文脈に大きく依存するため，正しく解析できるようにするためにはそれらを考慮した素性が必要となるだろう．

表 3: 複合辞に関する境界誤りの例

長単位列（正）	長単位列（誤）
損害賠償｜に｜関し｜,	損害賠償｜に関し｜,
進展如何｜に｜よる｜ところ	進展如何｜による｜ところ
選択する｜こととなり｜ます	選択する｜こと｜と｜なり｜ます

表 4: 品詞誤りの例

長単位	品詞（正）	品詞（誤）
高松市西部	名詞-固有名詞-地名-一般	名詞-普通名詞-一般
欧州連合	名詞-固有名詞-一般	名詞-普通名詞-一般
渡辺一男訳	名詞-普通名詞-一般	名詞-固有名詞-人名-一般

次に品詞誤りの傾向を調査した．名詞-普通名詞-一般と名詞-固有名詞-一般などの名詞同士の誤りが多く見られた．誤りの例を表 4 に示す．例えば，「欧州連合」を名詞-固有名詞-一般と判定すべきところを，名詞-普通名詞-一般と誤って判定した．これに対する改善策としては，固有名詞になりやすい名詞を学習データから取得し，素性として利用することが考えられる．

3.2.2 人が見たときに気になる誤りへの対処

品詞推定モデルでは，学習データにでてきたすべての品詞の中から最尤の品詞を推定するため，人間では誤らないような品詞が付与される可能性がでてくる．もし，BCCWJ の品詞体系で認められていない形式で品詞が付与されてしまうと，少数の誤りであっても人が見たときには目立つ誤りとなる．特に，研究対象になりやすく，自動解析も強く求められている複合辞などに対しては配慮する必要がある．例えば，「として」という助詞-格助詞の複合辞に対して，BCCWJ の品詞体系で認められていない助詞-係助詞などの品詞を

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html

[2] http://chasen.org/~taku/software/yamcha/

図 3: Comainu による長単位解析の実行例

付与したり，複合辞とは認められていない短単位列を複合辞として判定してしまうと，人が見たときに目立つ誤りとなり，解析器の信頼性が大きく低下してしまう．そのため，品詞推定モデルでは，複合辞については予め複合辞辞書を用意し，特定の品詞しか付与しないよう対処している．

4 長単位解析ツール Comainu

提案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．モデルの学習には BCCWJ のコアデータ（4,905 文，92,411 長単位，116,954 短単位）を用いた．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，Mecab[3] と Unidic 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 のいずれかを用いることができる．また，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おり，解析結果をファイルに保存することも可能である．

図 3 に Comainu による長単位解析の解析例を示す．図 3 の例では，平文を入力として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 2〜8 列目はそれぞれ短単位の書字形，発音系，語彙素読み，語彙素，品詞，活用型，活用形を表し，出力の 9〜14 列目はそれぞれ長単位の品詞，活用型，活用形，語彙素読み，語彙素，書字形を表す．

[3] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

5 まとめ

本論文では，長単位情報を自動付与する手法，及び，提案手法を実装した長単位解析システム Comainu について述べた．

本論文では，長単位解析の入力として適切な短単位列を想定したが，誤った短単位列が入力された場合にはもちろん長単位解析の誤りも増える．そのため，短単位解析の解析誤りが長単位解析に与える影響の調査，特に新たな言語単位や品詞体系を用いた場合にどのような影響がでるかを複数種類のコーパスを対象として比較調査することが今後必要となると考える．

参考文献

[1] Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the 6th International Language Resources and Evaluation*, pp. 1019–1024, 2008.

[2] Kiyotaka Uchimoto and Hitoshi Isahara. Morphological annotation of a large spontaneous speech corpus in Japanese. In *Proceedings of the 20th International Joint Conferences on Artificial Intelligence*, pp. 1731–1737, 2007.

[3] 小磯花絵, 小木曽智信, 小椋秀樹, 宮内佐夜香. コーパスに基づく多様なジャンルの文体比較-短単位情報に着目して. 言語処理学会第 15 回年次大会予稿集』, pp. 594–597, 2009.

[4] 近藤泰弘. 日本語通時コーパスの設計について. 国語研プロジェクトレビュー, Vol. 3, pp. 84–92, 2012.

[5] 田野村忠温. 日本語コーパスとコロケーション. 言語研究, Vol. 138, pp. 1–23, 2010.