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

短単位と長単位

語彙形態論研究に適した短単位と構文・意味研究に適した長単位
長単位はコーパスを用いた研究において重要な単位

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	。

長単位情報を自動付与する手法の提案
長単位解析：短単位列に対して，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

短単位						ラベル	長単位					
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	に関する	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	連用形-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		

長単位解析手法

1. チャンキングモデルにより長単位境界を認定．このとき，一部の長単位に対しては品詞情報も付与
2. カテゴリ推定モデルによって長単位の品詞，活用型，活用形を付与

チャンキングモデル

長単位の品詞情報なども認定するため，以下の4つのラベルの
尤もらしさを推定し，いずれかのラベルを付与

ラベル	長単位を構成する先頭の短単位であるか	長単位を構成する末尾の短単位で，かつ，品詞，活用型，活用形が長単位のものと一致するか
Ba	先頭	末尾かつ一致
Ia	先頭以外	末尾かつ一致
B	先頭	末尾かつ不一致／末尾以外
I	先頭以外	末尾かつ不一致／末尾以外

素性：着目する短単位と前後2短単位の以下の情報を利用

短単位情報
- 書字形，語彙素読み，語彙素，品詞，活用型，活用形，語種

汎化素性
- 階層化された素性に対して，上位階層で汎化した素性
  (例えば，「名詞-普通名詞-一般」に対して，「名詞」，「名詞-普通名詞」)

カテゴリ推定モデルによる後処理

BとIラベルのみから構成される長単位に対して，
尤もらしい品詞/活用型/活用形をそれぞれ推定

品詞推定モデル
- 学習データに現れた品詞の中から最尤の品詞を推定
- 助詞，助動詞かどうかは人手で整理した複合辞辞書とマッチング

固有	B	固有名詞	名詞-普通名詞-一般
名詞	Ia		
に	B	に関する	
関する	I		
論文	Ba	論文	名詞-普通名詞-一般

素性：着目する長単位と前後1長単位
各長単位を構成する先頭から2短単位と末尾から2短単位
- 書字形，語彙素読み，語彙素，品詞，活用型，活用形
- 汎化素性

評価実験

短単位情報は予め適切な情報が付与されていることを前提

データ：BCCWJのコアデータの一部を利用
学習データ 18,140 文（332,009長単位，419,414短単位）
テストデータ 2,015 文（ 36,297長単位， 45,906短単位）

モデル：
チャンキングモデル：CRF++
カテゴリ推定モデル：Yamcha (多項式カーネル，one-versus-rest)

ベースライン (Uchimoto et. al [2007])
- Ba, Iaは品詞，活用型，活用形が長単位のものと一致する短単位に付与
- チャンキングモデルの素性は短単位情報(語種を除く)のみ
- 自動獲得した書き換え規則による後処理

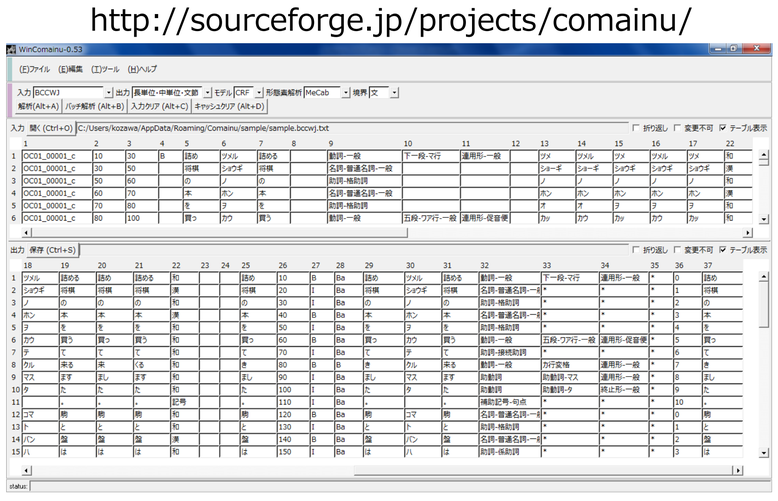
	境界認定			品詞認定		
	精度	再現率	F値	精度	再現率	F値
ベースライン	98.76%	98.76%	98.76	97.66%	97.71%	97.68
提案手法	98.94%	98.92%	98.93	98.67%	98.65%	98.66

境界誤り例	長単位列(正)	長単位列(誤)
名詞連続	医学部／倫理委員会	医学部倫理委員会
	春場所	春／場所
複合辞	損害賠償／に／関し／、	損害賠償／に関し／、
	選択する／こととなり／ます	選択する／こと／と／なり／ます

品詞誤り例	品詞(正)	品詞(誤)
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

長単位解析ツールComainu

http://sourceforge.jp/projects/comainu/



動作環境
- Perl: 5.10.0 以上
- Perl/Tk: 804.028 以上
- YamCha，CRF++

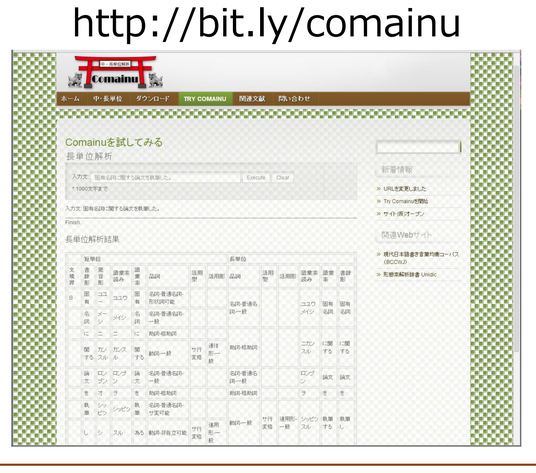
平文からの解析には以下も必要
- MeCab
- UniDic2

長単位解析 (Windows, Linux)
BCCWJのコアデータを用いて学習
（45,342文，828, 133長単位，1,047,069短単位）
入力：平文，または，短単位列（BCCWJ，KC）
出力：長単位を付与した短単位列を出力

長単位解析器の学習 (Linux)
短単位列を入力することにより，長単位解析モデル

その他
文節境界解析，中単位解析

http://bit.ly/comainu



今後の課題
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