CRF素性テンプレートの見直しによるモデルサイズを軽量化した解析用UniDic : unidic-cwj-2.2.0とunidic-csj-2.2.0

著者	岡 照晃
雑誌名	言語資源活用ワークショップ発表論文集
巻	2
ページ	144-153
発行年	2017
URL	http://doi.org/10.15084/00001515

CRF 素性テンプレートの見直しによる モデルサイズを軽量化した解析用 UniDic – unidic-cwj-2.2.0 と unidic-csj-2.2.0 –

岡 照晃 (国立国語研究所コーパス開発センター) *

UniDic for Morphological Analysis with reduced model size by review of CRF feature templates

Teruaki Oka (National Institute for Japanese Language and Linguistics)

要旨

国立国語研究所で構築している短単位自動解析用辞書『UniDic』は，現在，形態素解析器 MeCab 専用の解析用辞書として使用・公開を行なっている．しかし解析用 UniDic で使用している CRF の素性定義ファイルは，MeCab 用の他の辞書（『IPA 辞書』,『Juman 辞書』）に比べ，記述されているテンプレート数が多く，学習後の CRF の素性の重みベクトルの次元数も他辞書より大きい．そこで今回，現代語用の解析用『UniDic』の CRF テンプレートの再設計を行い，より少メモリかつ，これまでとほぼ同等の性能の解析用辞書を実現した．

1. はじめに

本稿では，最新版(1)の現代語の短単位自動解析(2)用辞書，書き言葉用の unidic-cwj-2.2.0 と，話し言葉用の unidic-csj-2.2.0 について述べる．UniDic は，国立国語研究所で構築・配布している国語研短単位 (小椋 2014) の電子化辞書 (伝ほか 2007) であり，①**設計方針**，②データベース（**UniDic データベース**），③形態素解析器 MeCab (Kudo et al. 2004) の解析用辞書（**解析用 UniDic**）を合わせた総称である（図 1）．本稿では，図 1③の解析用 UniDic を扱う．

解析用 UniDic の問題として，以前より，UniDic データベースに登録されている表層形（書字形出現形）数の増加に伴う MeCab 内部の CRF (Lafferty et al. 2001) モデルファイル(3)肥大化が指摘されている (鴻野知暁ほか 2014, Kono et al. 2015)．実際，既存の解析用 UniDic のモ

*teruaki-oka {at} ninjal.ac.jp

(1) 2017 年 9 月現在.

(2) UniDic データベースに登録されている言語単位（単位語）は，国語研の規定する短単位で揃えられており，解析用 UniDic を用いた MeCab の解析結果も，短単位の列となる．解析用 UniDic を使った形態素解析を本稿では**短単位自動解析**と呼ぶ.

(3) MeCab の CRF モデルファイルはヘッダー行を読み飛ばすと，CRF の素性の重みベクトルが 1 素性 1 行で記述されている（つまり，モデルファイルの行数≒素性の重みベクトルの次元数）．そのため本稿では，モデルファイルの行数を指して，**モデルサイズ**と呼ぶ.

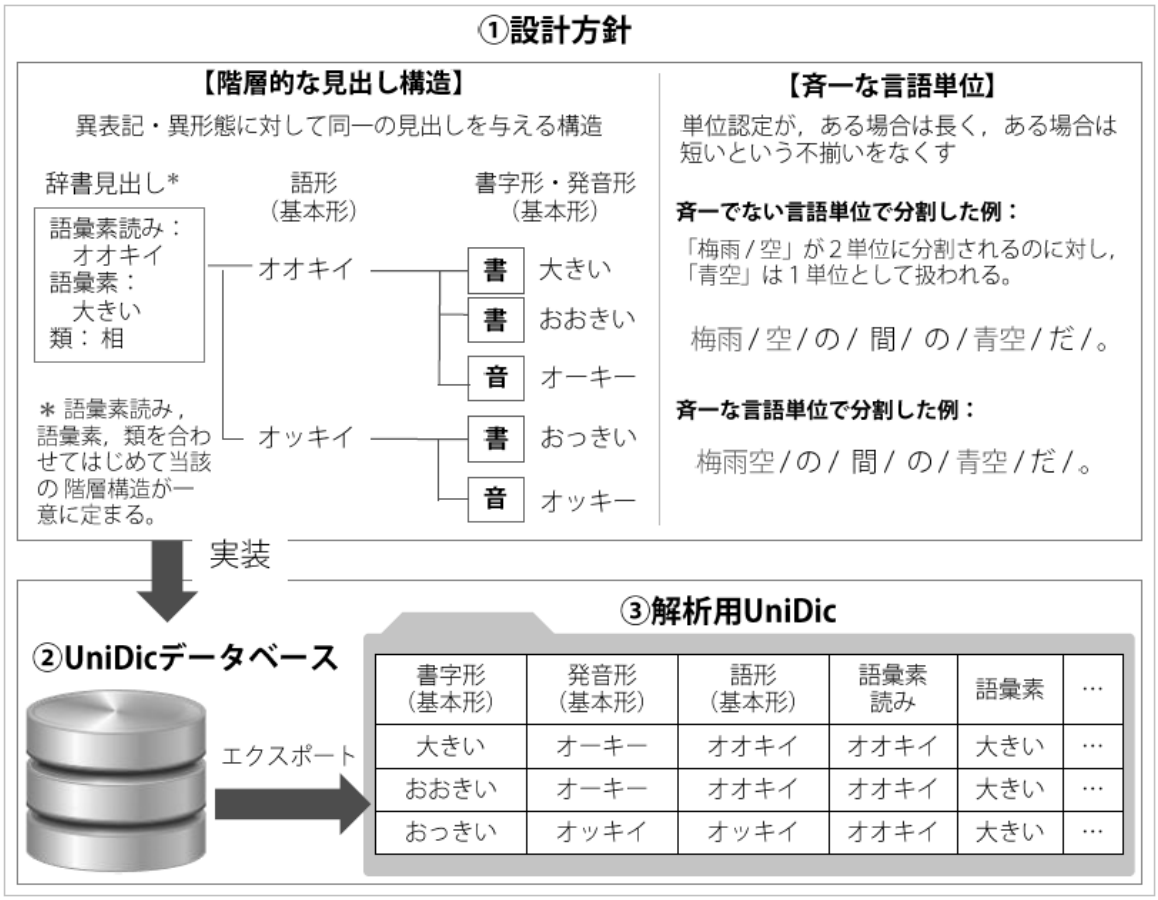

図 1　UniDic の全体像の概略.

表 1　MeCab の各解析用辞書のモデルサイズ比較．素性テンプレート数は，MeCab 用辞書の素性定義ファイル（feature.def）の空行とコメント行を削除した行数である（1 行 1 テンプレート）.

辞書	表層形数	CRF モデルサイズ（行数）	素性テンプレート数
unidic-mecab-2.1.2_src	756,463	6,758,363	263 (Unigram: 62, Bigram: 201)
mecab-ipadic-2.7.0	392,126	1,029,140	75 (Unigram: 20, Bigram: 55)
mecab-jumandic-7.0	751,185	573,020	56 (Unigram: 12, Bigram: 44)

デルファイル[4]と，他の MeCab の解析用辞書[5]ipadic[6]と jumandic[7]を表 1 のように比較すると，解析用 UniDic に登録されている表層形数は，jumandic と同等だが，モデルサイズはおよそ 10 倍である．各辞書で採用している品詞体系や，素性として使用可能な情報に差もあるが，解析用 UniDic の素性テンプレート数および，そこから展開される CRF のモデルファイルはあきらかに大きすぎる.

そこで，あらためて解析用 UniDic の素性定義ファイル（feature.def）を確認したところ，他

[4] unidic-mecab-2.1.2_src, unidic-mecab-2.1.2_model
[5] MeCab のサイト `http://taku910.github.io/mecab/` で公開されているもの.
[6] mecab-ipadic-2.7.0
[7] mecab-jumandic-7.0

の辞書に比べて，およそ 4〜5 倍の大きさで，さらに図 2 のような詳細な Bigram 素性が多量に定義されていることがわかった．これには，解析用 UniDic の素性定義ファイルを最初に作成した際，ipadic の素性定義ファイルをそのまま参考に解析用 UniDic の素性定義ファイルを作ったこと．また素性に語種情報を追加した際 (伝ほか 2008)，既存の素性テンプレート末尾に語種情報を追記したテンプレートを作り，そのまま元の素性定義ファイルに追加する，という方針をとったため，初期段階でも大きかった素性定義ファイルがさらに巨大なものとなった，という経緯がある．

本来，解析用 UniDic が付与する短単位は文脈に依存しない用例検索を目指して設計されている．そのため，解析用 UniDic も図 2 のような特定の連接に特化したような詳細な Bigarm 素性を作らずとも，汎用的に使える Bigram 素性だけに十分絞り込めるはずである．また UniDic の特徴の一つでもある階層的見出し構造も，現状では素性に十分に反映されているとはいえない．

そこで今回，階層的見出し語構造に基づく Unigram 素性と，統語的連接・語彙的連接に基づく Bigram 素性を中心に，解析用 UniDic の素性定義ファイルを再設計した．これにより，従来までと同等の解析性能を保ちつつ，CRF モデルサイズを縮小した現代書き言葉用の解析用 UniDic『unidic-cwj-2.2.0』を実現した．またそこからの追加学習で，話し言葉用の解析用 UniDic『unidic-csj-2.2.0』を作成した．

2. これまでの解析用 UniDic への CRF モデルサイズ縮小化の試み

解析用 UniDic のモデルサイズ肥大化の主な原因は，所内で整備しているコーパス（e.g., 通時コーパス (近藤泰弘 2012) など）の辞書を 1 つの UniDic データベースで管理していること．また，ある短単位の臨時的な表記に対応するための特殊な活用変化（特殊活用）が随時活用展開表へと追加されていることである．

(鴻野知暁ほか 2014, Kono et al. 2015) では，自至情報という，ある短単位がどの時代からどの時代までに使われていたかを表す情報を，語彙素，語形，書字形の段階ごとに付与する作業を UniDic データベースに対して行なった．自至情報を使うことで，通時コーパス構築用の各 UniDic（e.g., 中古和文 UniDic (小木曽智信ほか 2013) など）に登録する書字形出現形を絞り込むことができ，モデルサイズ縮小を実現した．また現在，国語研コーパス開発センター内では，増えすぎた特殊活用を削減する作業を始めている[8]．

これらはいずれも UniDic データベースを改善することで，結果的に解析用 UniDic のサイズを縮小しようとするアプローチである．対して本稿では，MeCab の学習時に使う素性テンプレートを再設計することで CRF モデルサイズを縮小する方法を採る．

3. 階層的見出し構造に着目した Unigram 素性の再設計

UniDic の特徴の一つに階層的見出し構造がある (伝ほか 2007)．階層の上から，語彙素，語形，書字形・発音形から成るが，図 3(a) のように書字形と発音形を同一階層とみなす場合もあ

[8] 2017 年 8 月現在.

図 2　unidic-mecab-2.1.2 の素性定義ファイル（feature.def）抜粋版.

```
～略～
# L[0]: pos1
# L[1]: pos2
# L[2]: pos3
# L[3]: pos4
# L[4]: cType
# L[5]: cForm
# L[6]: orth
# L[7]: orthBase
# L[8]: goshu
#
# R[0]: pos1
# R[1]: pos2
# R[2]: pos3
# R[3]: pos4
# R[4]: cType
# R[5]: cForm
# R[6]: orth
# R[7]: orthBase
# R[8]: goshu
～略～
BIGRAM GCW_GCW01:%L[0],%L?[4],%L?[5],%L?[8]/%R[0],%R?[4],%R?[5],%R?[8]
BIGRAM GCW_GCW02:%L[0],%L?[4],%L?[5],%L?[8]/%R[0],%R?[1],%R?[4],%R?[5],%R?[8]
BIGRAM GCW_GCW05:%L[0],%L?[1],%L?[4],%L?[5],%L?[8]/%R[0],%R?[4],%R?[5],%R?[8]
BIGRAM GCW_GCW06:%L[0],%L?[1],%L?[4],%L?[5],%L?[8]/%R[0],%R?[1],%R?[4],%R?[5],%R?[8]

BIGRAM OW_OW01:%L?[7],%L?[8]/%R?[7],%R?[8]

BIGRAM GOW_GOW01:%L[0],%L?[7],%L?[8]/%R[0],%R?[7],%R?[8]
BIGRAM GOW_GOW02:%L[0],%L?[7],%L?[8]/%R[0],%R?[1],%R?[7],%R?[8]
BIGRAM GOW_GOW03:%L[0],%L?[7],%L?[8]/%R[0],%R[1],%R?[2],%R?[7],%R?[8]
BIGRAM GOW_GOW04:%L[0],%L?[7],%L?[8]/%R[0],%R[1],%R[2],%R?[3],%R?[7],%R?[8]
BIGRAM GOW_GOW05:%L[0],%L?[1],%L?[7],%L?[8]/%R[0],%R?[7],%R?[8]
BIGRAM GOW_GOW06:%L[0],%L?[1],%L?[7],%L?[8]/%R[0],%R?[1],%R?[7],%R?[8]
BIGRAM GOW_GOW07:%L[0],%L?[1],%L?[7],%L?[8]/%R[0],%R[1],%R?[2],%R?[7],%R?[8]
BIGRAM GOW_GOW08:%L[0],%L?[1],%L?[7],%L?[8]/%R[0],%R[1],%R[2],%R?[3],%R?[7],%R?[8]
BIGRAM GOW_GOW09:%L[0],%L[1],%L?[2],%L?[7],%L?[8]/%R[0],%R?[7],%R?[8]
BIGRAM GOW_GOW10:%L[0],%L[1],%L?[2],%L?[7],%L?[8]/%R[0],%R?[1],%R?[7],%R?[8]
BIGRAM GOW_GOW11:%L[0],%L[1],%L?[2],%L?[7],%L?[8]/%R[0],%R[1],%R?[2],%R?[7],%R?[8]
BIGRAM GOW_GOW12:%L[0],%L[1],%L?[2],%L?[7],%L?[8]/%R[0],%R[1],%R[2],%R?[3],%R?[7],%R?[8]
BIGRAM GOW_GOW13:%L[0],%L[1],%L[2],%L?[3],%L?[7],%L?[8]/%R[0],%R?[7],%R?[8]
BIGRAM GOW_GOW14:%L[0],%L[1],%L[2],%L?[3],%L?[7],%L?[8]/%R[0],%R?[1],%R?[7],%R?[8]
BIGRAM GOW_GOW15:%L[0],%L[1],%L[2],%L?[3],%L?[7],%L?[8]/%R[0],%R[1],%R?[2],%R?[7],%R?[8]
BIGRAM GOW_GOW16:%L[0],%L[1],%L[2],%L?[3],%L?[7],%L?[8]/%R[0],%R[1],%R[2],%R?[3],%R?[7],%R?[8]
～略～
```

るが，図 3(b) のように書字形の下に発音形を設けて説明する場合もある．また，各階層は表 2 のような属性を持っている．

図 3(a)(b) と表 2 をまとめ，図 4，5 のような属性の階層構造を作成した．この図より，トッ

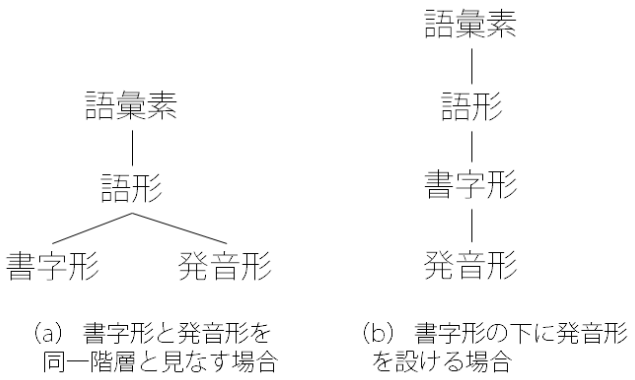

図 3　階層的見出し構造.

表 2　各階層に付与されている属性名

階層	属性
語彙素	語彙素（lemma), 語彙素読み (lForm), 語種 (goshu)
語形	語形基本形 (formBase), 語形出現形 (form), 品詞（大分類 (pos1), 中分類 (pos2), 小分類 (pos3), 細分類 (pos4)）, 活用型 (cType),　活用形 (cForm)
書字形	書字形基本形 (orth), 書字形出現形 (orthBase)
発音形	発音形基本形 (pron), 発音形出現形 (pronBase)

プダウンに属性を順次足していく操作で，Unigram 素性テンプレートを作っていく．図 4 は活用変化までを含めた素性を作るための構造であり，図 5 は活用変化を含めない汎化素性のための構造である．これにより，汎用的な素性（lemma+lForm のみ）から，より詳細化された（lemma+lForm～orth・orthBase・pron・pronBase）素性までが階層的に作成される．さらに【セイブツ-生物】【ナマモノ-生物】のように語彙素は異なる短単位同士の間でも汎化して学習が行えるよう，あえて語彙素を除き，語形から属性をテンプレートに追加していく階層的素性も使用する．

また図中にはないが，語種と文字種の Unigram 素性として，語種のみと，詳細化のための語種 + 品詞大分類，文字種のみと，詳細化のための文字種 + 品詞大分類の素性も設定した．その他，各属性のみの素性も用意している．

以上により，93 個の Unigram 素性テンプレートを作成した．

4.　統語的連接と語彙的連接に基づく Bigram 素性の再設計

統語的な連接は，品詞（とそれに準ずるもの）同士の連接と，品詞と活用形の連接である．品詞に準ずるものとは，以下のように UniDic データベース内で語彙化されている活用型と，CRF 学習時に語彙化される短単位である．

- 語彙化されている活用型: サ変変格-スル，形容詞-ナイ
- 語彙化される短単位: 助詞，助動詞，接辞

また，「ほり｜ごたつ」「一｜回（いっかい）」のような語頭語末音変化の連接素性もここに入る．

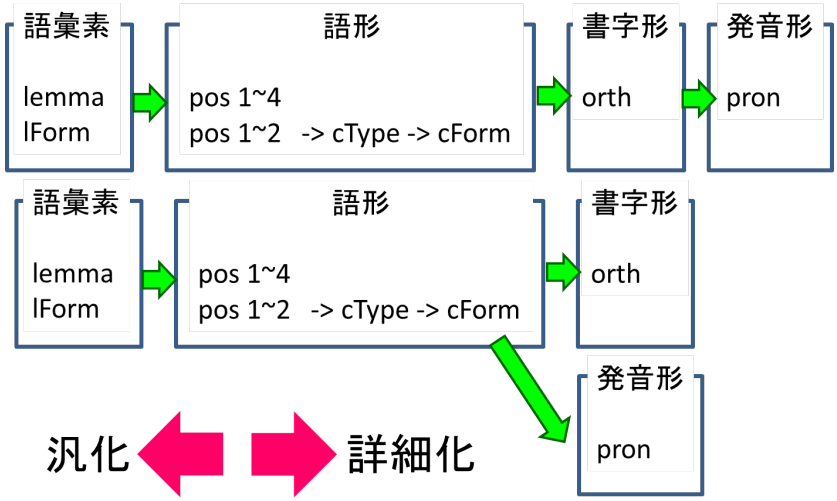

図 4　活用変化を考慮した階層的 Unigram 素性を作るための構造．上段が図 3(b)，下段が図 3(a) に対応している．

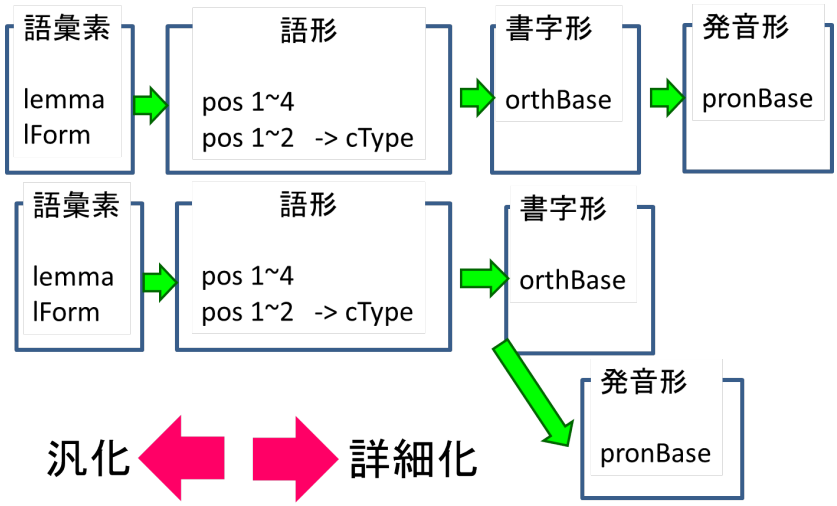

図 5　活用変化を考慮しない汎化用の階層的 Unigram 素性を作るための構造．上段が図 3(b)，下段が図 3(a) に対応している．

語彙的連接は，ipadic や jumandic の leaf 素性にあたる品詞 + 活用の連接素性に，語種同士，活用型同士，語種と活用型の連接加えたものである．

以上により，81 個の Bigram 素性テンプレートを作成した．

5.　再設計した素性定義ファイルの性能評価実験

再設計した素性定義ファイルの有効性を確認するため，現代語の unidic-2.1.2 で使用している素性定義ファイルとの比較を行なった．

表 3　学習用コーパス

コーパス名	文字数	短単位数	文数
ALL_TRAIN	2,941,919	1,867,570	80,888
SP_TRAIN	1,566,518	988,179	44,777

表 4　評価用コーパス

コーパス名	文字数	短単位数	文数
ALL_TEST	3,243,03	205,577	8,991
SP_TEST	174,289	109,548	4,977

5.1　実験設定

解析用辞書のエントリはすべての辞書で共通とし, 文献 (鳴野知暁ほか 2014, Kono et al. 2015) の時代情報を使った絞り込みによって UniDic データベース (小木曽智信ほか 2014) から取り出した 860,073 の表層形（辞書のキー）がエントリされている．コーパスは，BCCWJ と CSJ のコアデータを使い (ALL)，BCCWJ の中でも話し言葉に近い（PB，OC，OY）と CSJ を部分コーパスとして使用する（SP）．BCCWJ は各ジャンルで，CSJ は全体を文単位にそれぞれ 9:1 で学習・評価にわけた．内訳を表 3，表 4 に示す．

MeCab のバーションは 0.996 を使用し，学習時の引数は，並列化オプションを除きすべてデフォルトのまま使用した．また CRF の正規化項のハイパーパラメータ C (=σ^2) も，全学習において共通にデフォルトの 1.0 に設定した．

評価は文献 (小木曽智信ほか 2013, 小木曽智信ほか 2014) と同じく，境界認定，品詞認定，語彙素認定，発音認定の 4 段階で F1 値を評価する．日本語の自動形態素解析における F1 値の計算方法は文献 (Kudo et al. 2004) を参照．境界認定は，文中での開始位置と終了位置が両方正しく認定できた（正解データと一致した）短単位数を評価している．品詞認定は，境界認定をパスした短単位の内，品詞大分類，中分類，小分類，細分類，活用型，活用形がすべて正しく認定できた短単位数を評価している．語彙素認定は，品詞認定をパスした短単位の内，語彙素読み，語彙素の両方が正しく認定できた短単位数を評価している．発音認定は，語彙素認定をパスした短単位の内，発音形出現形が正しく認定できた短単位数を評価している．評価には形態素解析器性能評価ツール MevAL[(9)]の Beta 版を使用した．

また本実験では，ALL_TRAIN で学習した unidic-cwj と，unidic-cwj のモデルファイル（model）に SP_TRAIN で MeCab の**追加学習（Regularized Adaptation**[(10)]）した unidic-csj を作成し，比較を行なった．

(9) https://teru-oka-1933.github.io/meval/

(10) 初期のモデルパラメータをできるだけ変更せずに，新しい学習データにできるだけ適応するような新しいモデルを学習する手法．MeCab のサイトでは「再学習」と書かれている．MeCab の-M オプション（初期モデルの指定）で使用可能．詳細は文献 (Imamura 2013) を参照．

表 5　unidic-cwj における短単位自動解析性能の評価.

評価用コーパス	未知語	素性テンプレート	境界認定	品詞認定	語彙素認定	発音認定
ALL	なし	従来 (2.1.2)	99.42	98.20	97.86	97.21
		再設計 (2.2.0)	99.42	98.17	97.89	97.31
	書字形レベル	従来 (2.1.2)	98.21	96.65	96.23	95.58
		再設計 (2.2.0)	98.23	96.63	96.27	95.68
	語彙素レベル	従来 (2.1.2)	98.54	97.19	96.81	96.17
		再設計 (2.2.0)	98.54	97.18	96.84	96.28
SP	なし	従来 (2.1.2)	99.16	97.46	97.11	96.42
		再設計 (2.2.0)	99.16	97.43	97.12	96.46
	書字形レベル	従来 (2.1.2)	97.99	95.95	95.52	94.82
		再設計 (2.2.0)	98.00	95.91	95.53	94.87
	語彙素レベル	従来 (2.1.2)	98.35	96.54	96.14	95.45
		再設計 (2.2.0)	98.35	96.51	96.15	95.51

表 6　unidic-csj における短単位自動解析性能の評価.

評価用コーパス	未知語	素性テンプレート	境界認定	品詞認定	語彙素認定	発音認定
ALL	なし	従来 (2.1.2)	99.36	97.97	97.56	96.86
		再設計 (2.2.0)	99.34	97.96	97.61	96.94
	書字形レベル	従来 (2.1.2)	98.41	96.74	96.27	95.55
		再設計 (2.2.0)	98.40	96.72	96.31	95.63
	語彙素レベル	従来 (2.1.2)	98.60	97.09	96.63	95.92
		再設計 (2.2.0)	98.60	97.09	96.70	96.03
SP	なし	従来 (2.1.2)	99.16	97.47	97.13	96.46
		再設計 (2.2.0)	99.14	97.47	97.16	96.53
	書字形レベル	従来 (2.1.2)	97.54	95.37	94.91	94.22
		再設計 (2.2.0)	97.54	95.38	94.94	94.28
	語彙素レベル	従来 (2.1.2)	97.91	96.04	95.61	94.94
		再設計 (2.2.0)	97.92	96.07	95.67	95.04

6. 実験結果

表 5 と表 6 に結果を示す．未知語列は，「なし」の場合，UniDic データベースからダンプした書字形出現形をすべて使用している．「書字形レベル」の場合，評価用コーパスでしか出現しない書字形出現形を解析用 UniDic の seed の csv からすべて除外して学習を行なっている．「語彙素レベル」の場合，評価用コーパスでしか出現しない書字形出現形を，階層的見出し構造を利用し，当該書字形の祖父にあたる語彙素の下の書字形出現形を解析用 UniDic の seed の csv からすべて除外して学習を行なっている．

2 つの結果を見比べると，素性テンプレート書き換えてもほぼ同等の性能が確保されていることがわかる．ただし，フルサイズのコーパス（TRAIN+TEST）で学習した場合，モデルサイズが 5,058,020 行と，2.1.2 からおよそ 170 万行の削減ができた．品詞認定の性能が若干低くなっているが，エラーを見ると，助動詞「に」と格助詞「に」の誤りが増加していた．ただし，助動詞「に」と格助詞「に」の区別は，短単位の連接だけで区別するのがもともと難しく，従来のものでは細かく作った Bigram 素性でオーバーフィットたと考えられる．また性能的にも品詞認定で 2.1.2 のほうに有意差は見られなかった．

また，unidic-csj は ALL_TRAIN で追加学習を行なったが，SP_TEST でも unidic-cwj が同等の性能だった．ただし，未知語がほとんどないような状態では unidic-csj の方が，若干高い性能となっている．

またいずれの結果でも「語彙素レベルでの未知語作成」した場合の方が，「書字形レベルでの未知語作成」した場合よりも性能が高くなっているが，これは書字形レベルで未知語を作成しただけでは，同一語彙素の別の語として解析される場合が多く，その場合，誤って選ばれた語は正解の表記よりも短い表記で，そこで過分割が生じたためである．

7. おわりに

本稿では，MeCab の素性定義ファイルの再設計により，従来と同等の性能を保ちつつ，CRF モデルサイズの縮小化を実現した．また，BCCWJ と CSJ のコアデータをすべて学習に使った unidic-cwj-2.2.0 と，そのモデルファイルに話し言葉寄りの部分コーパスで追加学習した unidic-csj-2.2.0 を作成した．全体的な性能では，unidic-cwj の方が高いものの，未知語が少ない状況ならば unidic-csj のほうが性能が高くなることも分かった．

今後は，更なる素性定義ファイルの改良や，rewrite.def の改善により，解析用 UniDic の更なる性能向上と，少メモリ化に引き続き取り組んでいく．

謝　辞

本研究は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021 年度) の成果である。

文　献

[Imamura 2013] Kenji Imamura (2013). “Case Study of Model Adaptation: Transfer Learning and Online Learning.” *Proceedings of IJCNLP-2013 (the 6th International Joint Conference on Natural Language Processing)*, pp. 1292–1298.

[Kudo et al. 2004] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto (2004). “Applying Conditional Random Fields to Japanese Morphological Analysis.” *Proceedings of EMNLP-2004 (the 2004 Conference on Empirical Methods in Natural Language Processing)*, pp. 230–237.

[Kono et al. 2015] Tomoaki Kono and Toshinobu Ogiso (2015). “Improving an Electronic Dictionary for Morphological Analysis of Japanese: Use of historical period information.” *Proceedings of ICML-2001 (the 18th International Conference on Machine Learning)*, pp. 282–289.

[Lafferty et al. 2001] John Lafferty, Andrew McCallum and Fernando Pereira (2001). “Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” *Proceedings of ICML-2001 (the 18th International Conference on Machine Learning)*, pp. 282–289.

[小木曽智信ほか 2013] 小木曽智信・小町守・松本裕治 (2013).「歴史的資料を対象とした形態素解析」自然言語処理，20:5, pp. 727–748.

[小木曽智信ほか 2014] 小木曽智信・中村壮範 (2014).「『現代日本語書き言葉均衡コーパス』

形態論情報アノテーションシステムの設計・実装・運用」自然言語処理，21:2, pp. 301-332.

[小椋 2014] 小椋 秀樹 (2014).「形態論情報」講座日本語コーパス 2 巻 『書き言葉コーパス 設計と構築』, 4 章, pp. 68–88.

[鴻野知暁ほか 2014] 鴻野知暁・小木曽智信 (2014).「見出し語の時代情報を付与した電子化辞書の構築」言語処理学会 第 20 回 年次大会 発表論文集，pp. 209–212.

[近藤泰弘 2012] 近藤泰弘 (2012).「日本語通時コーパスの設計」NINJAL「通時コーパス」プロジェクト・Oxford VSARPS プロジェクト合同シンポジウム「通時コーパスと日本語史研究」予稿集，pp.1–10.

[伝ほか 2002] 伝康晴・宇津呂武仁・山田篤・浅原正幸・松本裕治 (2002).「話し言葉研究に適した電子化辞書の設計」第 2 回「話し言葉の科学と工学」ワークショップ講演予稿集，pp. 39–46.

[伝ほか 2007] 伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元 清貴・小磯 花絵 (2007).「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』，22 号，pp.101–123.

[伝ほか 2008] 伝康晴・中村純平・小木曽智信・小椋秀樹 (2008).「語種情報を用いた同表記異音語の解消」言語処理学会 第 14 回 年次大会 発表論文集，pp. 69–72.