unidic-mecab ver. 2.1.2 ユーザーズマニュアル

The UniDic Consortium

2013 年 1 月

unidic-mecab ver. 2.1.2 Users Manual
The UniDic Consortium

ver. 1.3.0 2 April 2007
ver. 1.3.5 12 October 2007
ver. 1.3.8 25 April 2008
ver. 1.3.9 15 July 2008
ver. 1.3.11 30 April 2009
ver. 1.3.12 31 July 2009
ver. 2.1.0 27 February 2011
ver. 2.1.1 12 December 2012
ver. 2.1.2 26 January 2013

目次

はじめに

UniDic は、形態素解析システム用の日本語辞書です。MeCab 用（unidic-mecab）があります。MeCab は言語処理のためのソフトウェアとしてフリーで公開され、広く用いられています。

UniDic は、従来の形態素解析辞書と比べ、言語学・国語学や音声情報処理など、より多様な目的に適した体系にもとづくものです。具体的には、以下の特徴を持ちます。

- 国立国語研究所で規定した「短単位」という揺れがない斉一な単位で設計されています。
- 語彙素・語形・書字形・発音形の階層構造を持ち，表記の揺れや語形の変異にかかわらず同一の見出しを与えることができます。
- アクセントや音変化の情報を付与することができ、テキスト音声合成などに利用することができます。

UniDic の開発には、情報処理振興事業協会「擬人化音声対話エージェント基本ソフトウェアの開発」プロジェクト（代表：東京大学・嵯峨山茂樹）、情報処理学会「音声対話技術コンソーシアム」（ISTC）（代表：豊橋技術科学大学・新田恒雄）、文部科学省科学研究費補助金特定領域研究「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」（平成 18～22 年度、領域代表者：国立国語研究所・前川喜久雄）からの助成を得ています。また平成 18 年度からは、国立国語研究所の研究課題「大規模汎用日本語データベースの構築とその活用に関する調査研究」のもと、同研究所研究開発部門言語資源グループ（現言語資源研究系・コーパス開発センター）と共同開発を行なっています。

本辞書に関するお問い合わせは以下にお願いします。

Tel: 042-540-4300（国立国語研究所代表）
E-mail: unidic@ninjal.ac.jp

1 unidic-mecab とは

本文書では unidic-mecab について説明します。unidic-mecab は形態素解析システム MeCab 用の辞書であり、UniDic2 形態論辞書の派生物の 1 つです。形態論辞書そのものの説明については UniDic2 形態論辞書マニュアルを、形態素解析辞書の生成および拡張については UniDic Tools マニュアルを参照してください。

2 インストール

unidic-mecab のインストールおよび実行には、MeCab（ver. 0.96 以降）[*1]が必要です。あらかじめインストールしておいてください。

2.1 パッケージ版のインストール

パッケージ版（Windows 用）をお使いの方は、パッケージをインストールすればそのまま『茶まめ』からお使いいただけます。

1. 圧縮ファイル unidic-mecab-2.1.2_windows.zip を解凍します。
2. INSTALL.exe を実行します。辞書は C:\Program Files\unidic フォルダ内の dic\unidic-mecab フォルダの下にインストールされます。

2.2 バイナリ辞書のインストール

バイナリ辞書をインストールする場合は以下の手順に従ってください。

1. 圧縮ファイル unidic-mecab-2.1.2_bin.zip を解凍します。
2. ステップ 1 でできたフォルダを MeCab 辞書フォルダに移動します。

2.3 ソース辞書のインストール

ソース辞書をインストールする場合は以下の手順に従ってください。

1. 圧縮ファイル unidic-mecab-2.1.2_src.zip を解凍します。
2. ステップ 1 でできたディレクトリに移動し、以下のコマンドを実行します。

```
./configure && make
```

3. 管理者権限で以下のコマンドを実行します。

```
make install
```

辞書は標準では/usr/local/lib/mecab/dic/unidic にインストールされます。

[*1] http://mecab.sourceforge.net/

3 unidic-mecab のファイル群

3.1 語彙定義ファイル

lex.csv には語彙項目の一覧が記述されています。各語彙項目には、UniDic2 形態論辞書の基本属性がカンマ区切りで記載されています。

1. 品詞大分類
2. 品詞中分類
3. 品詞小分類
4. 品詞細分類
5. 活用型
6. 活用形
7. 語彙素読み
8. 語彙素（語彙素表記 + 語彙素細分類）
9. 書字形出現形
10. 発音形出現形
11. 書字形基本形
12. 発音形基本形
13. 語種
14. 語頭変化型
15. 語頭変化形
16. 語末変化型
17. 語末変化形

これらの属性の詳細については、UniDic2 形態論辞書マニュアルを参照してください。また、これら以外の属性を UniDic2 辞書データベースから取得し、形態素解析済みテキストに付加することもできます[*2]。詳しくは UniDic Tools マニュアルを参照してください。

3.2 その他の定義ファイル

.def の拡張子を持つファイルには、連接規則など、各種のモデル定義情報が記述されています。詳しくは http://mecab.sourceforge.net/ の解説を参照してください。

3.3 dicrc ファイル

dicrc ファイルには、MeCab の実行に必要なさまざまなオプションが定義されています。以下では、形態素結果出力に関するものを説明します。

[*2] 自然言語処理でよく用いられる仮名形や音声合成で用いられるアクセント型などをあらかじめ lex.csv に追記した別パッケージもあります。

```
output-format-type = unidic

node-format-unidic = %m\t%f[9]\t%f[6]\t%f[7]\t%F-[0,1,2,3]\t%f[4]\t%f[5]\n
unk-format-unidic  = %m\t%m\t%m\t%m\t%F-[0,1,2,3]\t%f[4]\t%f[5]\n
bos-format-unidic  =
eos-format-unidic  = EOS\n
```

出力フォーマットを変更するには、output-format-type に任意のフォーマット名 XXX を記載し、node-format-XXX, unk-format-XXX, bos-format-XXX, eos-format-XXX でそれぞれ単語・未知語・文頭・文末の出力形式を指定します。なお、f[N] で指定される各属性の意味については、dicrc ファイルの冒頭にあるコメントを参照のこと。

dicrc のその他のオプションについては、http://mecab.sourceforge.net/ の解説を参照してください。

付録 A　変更履歴

A.1　Ver. 1.3.0 → ver. 1.3.5

- 活用のある語の「基本形」を「終止形」と「連体形」に区別

A.2　Ver. 1.3.5 → ver. 1.3.8

- MeCab 版を作成
- 語種情報を追加

A.3　Ver. 1.3.8 → ver. 1.3.9

- MeCab 版の辞書の並び順を変更

A.4　Ver. 1.3.9 → ver. 1.3.11

- 「補助記号-ＡＡ-{ 顔文字, 一般 }」を新設

A.5　Ver. 1.3.11 → ver. 1.3.12

- 「名詞-固有名詞-組織名」を廃止
- MeCab 版の辞書に仮名形やアクセント型などを出力

A.6　Ver. 1.3.12 → ver. 2.1.0

- UniDic2 の辞書データベースから生成するように変更
- 「名詞-普通名詞-助数詞可能」を新設
- 語彙素や品詞分類を整理
- MeCab 版に一本化

A.7　Ver. 2.1.0 → ver. 2.1.1

- MeCab-0.994 に対応

A.8　Ver. 2.1.1 → ver. 2.1.2

- 仮名形やアクセント型を追記したパッケージを追加