UniDic version 2.1.1 Windows 版パッケージ

「茶まめ」マニュアル

2012 年 12 月 13 日　小木曽 智信

1. はじめに

「茶まめ」は UniDic を使って形態素解析を行うのを補助するためのソフトウェアです。茶まめを使うことにより、UniDic で解析する際に必要な一連の作業を、わかりやすいインターフェイス（GUI）で行うことができます。このマニュアルでは茶まめの使い方について説明します。

※ UniDic 2.x 以降、解析器は MeCab のみに対応しています。ChaSen への対応は中止し、XSLT による解析後処理（ChaOne）もなくなりました。

※ 茶まめを使わなくても UniDic による解析を行うことができます。このマニュアルの「0. コマンドプロンプトでの利用」をご覧ください。

2. 茶まめの使い方

2.1. 起動

デスクトップに作られる「茶まめ」アイコン、または、スタートメニューの「UniDic」→「茶まめ」から起動してください。

次の画面が現れます。

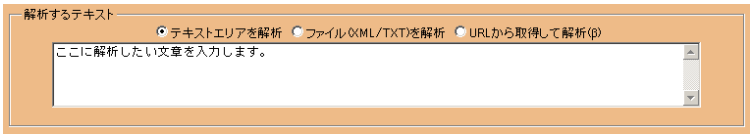
この画面の上から下へと順に処理方法を指定してゆき、最後に「実行」ボタンを押すことで結果を出力します。

解析器として「MeCab」がインストールされていないとこの時点で警告が出ます。インストールマニュアルを参考にして、インストールしてください。

2.2. 解析対象テキストの設定

画面上部の「解析するテキスト」で解析対象のテキストを指定します。解析対象をラジオボタンで選択してください。選択した方法にあわせて画面が変わります。

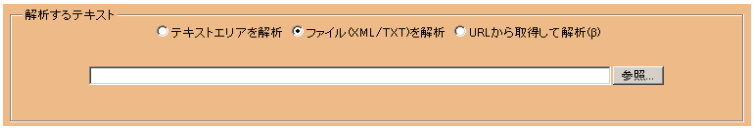
- テキストを入力したり貼り付けたりする場合には、「テキストエリアを解析」を選びます（起動時にはこれが選ばれています）。テキストエリア（白い部分）にテキストを入力してください。テキストエリアをダブルクリックすると内容がクリアされます。



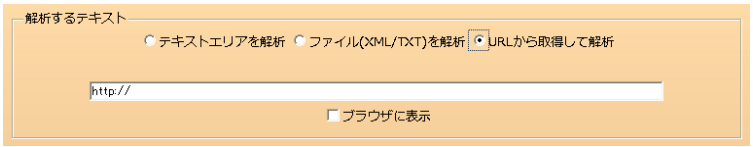
- 解析対象のファイルを指定する場合には、「ファイル(XML/TXT)を解析」を選びます。その後「参照」ボタンを押してファイルを指定してください。指定できるファイルはテキストファイル・XML ファイル・HTML ファイルです。HTML ファイルはタグを除去してテキストとして解析します。テキストファイルの文字コードは自動判別されます（Shift_JIS, EUC-JP, JIS(ISO-2002-JP), UTF-8, UTF-16）。

  ※ワイルドカードを使って複数のファイルを指定し、一度に処理することができます。

  例：C:¥DATA¥*.txt → C:¥DATA の中の全ての txt ファイルを解析します



- インターネット上からダウンロードして解析する場合には、「URL から取得して解析 」を選んで URL を指定してください。「ブラウザに表示」オプションをチェックすると、解析後にブラウザで当該ページを表示します。

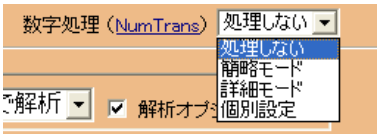


2.3. 解析前処理の設定

必要に応じてチェックボックスをチェック（☑）して解析前処理を指定します。

- 「☐ 半角文字を全角に変換」をチェックすると解析前に文字を変換します。UniDic は原則として全角文字にしか対応していません。半角文字列が入ったテキストを解析する場合にはこれをチェックしてください。

- 「数字処理（NumTrans）」の右側にあるセレクトボックスで指定することで、解析前に数字を解析しやすい形に変換することができます。変換の仕方として、「簡易モード」「詳細モード」と「個別設定」（設定ファイルを利用する場合）が選べます。詳細については NumTrans のマニュアルをご覧ください。標準では「処理しない」になっています。



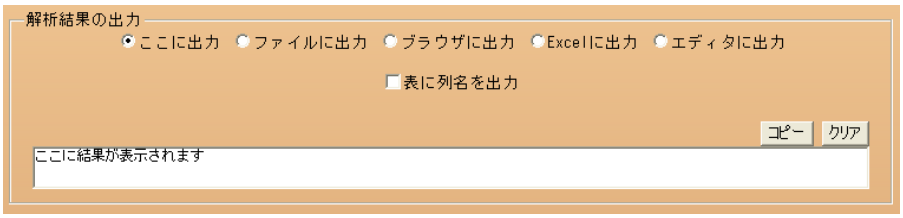
2.4. 解析の設定

解析器とは形態素解析を行うプログラムのことです。「茶まめ」は現在、「MeCab」のみに対応しています。「解析しない」というオプションがありますが、これは解析を行わないで、前処理だけを行った結果を出力する場合に使用します。

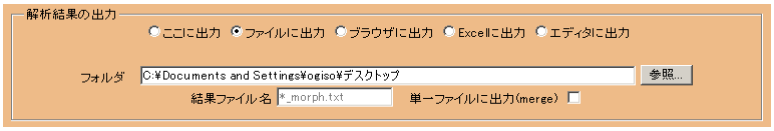
- MeCab の辞書を変えたい場合には、ここで辞書があるフォルダを指定します。
- Windows 版 MeCab 付属の IPADIC の場所（C:¥Program Files¥Mecab¥doc¥ipadic）を指定すると、UniDic ではなく IPADIC で解析することができます。

2.5. 解析結果の出力

画面下部の「解析結果の出力」で解析結果の出力先を指定します。ラジオボタンで出力の方法を選択してください。選択した方法にあわせて画面が変わります。



- 「ここに出力」を選択すると、茶まめのテキストエリア（画面下部）に結果を表示します。短い文章を解析する場合に適しています。
- 「ファイルに出力」を選択すると、出力先の指定画面が現れ、指定したファイルに解析結果を出力することができます。比較的大きなファイルを処理する場合にはファイルに出力してください。



ワイルドカードを使って入力ファイルを指定した場合には「ファイルに出力」ボタンしか選択できません。また、ワイルドカードを利用する場合に限って「単一ファイルに出力」チェックボックスをチェック（☑）することで解析結果を一つのファイルにまとめることができます。

- 「ブラウザに出力」を選択すると、解析結果を Web ブラウザ（Internet Explorer）に

出力します。解析結果の XML ファイルを閲覧する場合に利用してください。

- 「Excel に出力」を選択すると、解析結果を Microsoft Excel に出力します。解析結果を表形式テキストに変換して表として使いたい場合に利用してください。
- 「エディタに出力」を選択すると、解析結果をテキストエディタに出力します。選択されるエディタは、Internet Explorer の「ソースの表示」用に指定されているエディタです。
- 「表に列名を出力」をチェックすると、表形式で出力するときの先頭行に列名を出力します。

3. 茶まめの出力形式

茶まめが出力するファイルの形式について説明します。

3.1. ファイルの文字コード

解析結果は辞書と同じ UTF-8 で出力されます。

3.2. 表形式テキストのフィールド

標準状態で出力される表形式テキスト（xml2txt によって変換したタブ区切りテキスト）のフィールドは左から順に次の通りです。

出典　文境界　書字形　発音形　語彙素読み 語彙素　品詞　活用型　活用形　語形　書字形基本形　語種

※文境界は B が文頭、I がそれ以外。

表形式テキスト出力の例

出典	文境界	書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	書字形基本形	語種
chamame	B	ここ	ココ	ココ	此処	代名詞			ココ	ここ	和
chamame	I	に	ニ	ニ	に	助詞-格助詞			ニ	に	和
chamame	I	解析	カイセキ	カイセキ	解析	名詞-普通名詞-サ変可能			カイセキ	解析	漢
chamame	I	し	シ	スル	為る	動詞-非自立可能	サ行変格	連用形-一般	スル	する	和
chamame	I	たい	タイ	タイ	たい	助動詞	助動詞-タイ	連体形-一般	タイ	たい	和
chamame	I	文章	ブンショー	ブンショウ	文章	名詞-普通名詞-一般			ブンショウ	文章	漢
chamame	I	を	オ	ヲ	を	助詞-格助詞			ヲ	を	和
chamame	I	入力	ニューリョク	ニュウリョク	入力	名詞-普通名詞-サ変可能			ニュウリョク	入力	漢
chamame	I	し	シ	スル	為る	動詞-非自立可能	サ行変格	連用形-一般	スル	する	和
chamame	I	ます	マス	マス	ます	助動詞	助動詞-マス	終止形-一般	マス	ます	和
chamame	I	。			。	補助記号-句点				。	記号

なお、解析対象テキストの中の半角スペースは解析時に消去されます。

3.3. 出力される XML のタグ

標準の XML 形式による出力では、茶まめの処理により次のようなタグ・属性が付与されます。すべて名前空間 URI として「http://www.unidic.org/chasen/ns/structure/1.0」を使用します。名前空間接頭辞として「cha:」を使用しています。

要素・属性	説明
cha:D	テキスト・HTML ファイルを処理した場合に、XML 文書のルートタグとして使われます。
cha:S	W1 を含み込む文レベルの要素です。テキスト・HTML ファイルの場合には EOS から BOS の範囲に相当します。XML ファイルの場合には解析後に cha:W1 の親として挿入します。
cha:W1	解析された短単位のタグです。元のテキストを内容として、次の属性が付けられます（chasenrc による解析の場合）。 • orth：書字形 • kana：仮名形 • pron：発音形 • pos：品詞 • cType：活用型 • cForm：活用形 • orthBase：書字形基本形 • kanaBase：仮名形基本形 • pronBase：発音形基本形 • lForm：語彙素読み • lemma：語彙素表記 • form：語形 • aType：アクセント型 • aConType：アクセント結合型 • goshu：語種
@cha:src	XML 文書のルートタグに付けられる属性で、解析対象の出典を表します。入力がテキストエリアの場合は「chamame」、ファイルの場合はファイル名、URL 指定の場合は URL が値となります。 （グローバル属性で名前空間接頭辞がつきます）

NumTrans によって処理を行った場合にはこれ以外のタグも出力されます。それらのタグについては NumTrans のマニュアルをご覧ください。

4. 茶まめが行う処理

※ この項目は茶まめが内部的に行っている処理について説明したものです。一般的な利用を行うだけであれば読み飛ばしてかまいません。

4.1. 茶まめの処理の流れ

茶まめは次のような流れで処理を行います。

1 入力
2 解析対象の XML 文書化
3 解析前処理 XSLT
  3.1 半角文字の変換（オプション）
  3.2 NumTrans（オプション）
  3.3 タグの除去（MeCab 利用時のみ）
4 解析器（MeCab）による解析
5 出力

4.2. 解析対象の XML 文書化処理

UniDic 関連ツールは入出力を XML で行います。そのため、テキストエリアに入力された文字列なども XML 文書に変換してから解析前処理へ進みます。この XML 文書化の処理方法は、次の表の通りです。

入力の種類		処理方法
テキストエリア		テキストファイルと同じ処理を行う。
ファイル	テキスト	ルートを cha:D タグで囲み、「。」を区切りとして cha:S タグを挿入して XML 文書にする。特殊文字（<>&）は全角文字に置き換える。 文字コードは次の各種エンコーディングを自動判別する。 （Shift_JIS, EUC-JP, JIS(ISO-2002-JP), UTF-8, UTF-16）
	HMTL	タグを除去（br, tr, li, p などは改行を挿入）してテキスト化したのち、テキストファイルと同様に XML 文書にする。 実体参照や特殊文字は全角文字や = に置き換える。
	XML	ルートにネームスペース宣言、cha:src 属性を付加する。文書型宣言付きの場合は削除する。 （解析後、Inserts.xsl により cha:S タグを挿入する）
URL 指定		データを取得してファイルに保存後、種類別に処理を行う。 （ただし XML ファイルとして処理するのは拡張子 XML, RDF のファイルのみ）

5. コマンドプロンプトでの利用

きわめて大きなファイルを解析する場合や、独自の処理を行いたい場合などには、コマンドプロンプト上で解析を行ってください。

MeCab の場合

MeCab の解析用辞書として UniDic（UniDic-mecab）を使う場合、MeCab の-d オプションで unidic-mecab フォルダを指定することにより、UniDic を使った解析を行うことができます。

インストール直後の状態では、コマンドプロンプトで次のように入力することで UniDic による解析が可能です。

```
"C:¥Program Files¥mecab¥bin¥mecab.exe" -d "C:¥Program Files¥unidic¥dic¥unidic-mecab" < 《入力ファイル》 > 《出力ファイル》
```

※ 入力ファイルの文字コードは UTF-8 にしておく必要があります。またインストール先を変更した場合にはそれにあわせてください。

※ 実際に利用する場合には mecab のインストールパスを環境変数 Path に追加するなどして環境を整備してください。

なお、XSLT による変換をコマンドプロンプト上で行いたい場合には、msxsl.exe*などのコマンドラインツールを利用してください。

6. FAQ（よくある質問）

Q：大きなテキストを解析したいのですが、どのくらいのサイズまで解析できますか。

A：**解析器**（MeCab）自体は数百 MB 程度のファイルでも解析することができます。しかし XSLT による前処理・後処理を行う場合には多くのメモリを必要としますので、数 MB 程度のテキストでも変換できない場合があります。

茶まめでは XSLT を多用するため、標準で 10MB 以上のファイルは処理しないようにしています。また、一つの入力ファイルの解析結果が 50MB を越えると XSLT の後処理をパスするようになっています（これらの数字は「茶まめ.hta」ファイルを直接書き換えることで変更できます)。

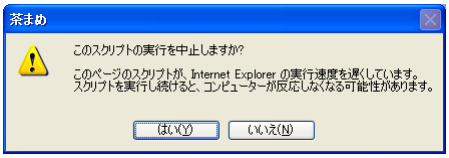
茶まめを使って大きなデータを解析する場合には、小さめのファイルに分割して処理してください。茶まめでは、ワイルドカードを使って複数のファイルをまとめて解析することができるようになっています。また、「単一ファイルに出力（merge)」をチェックすることで、出力結果を一つのファイルにまとめることもできます。

どうしても大きなファイルを処理する必要がある場合は、コマンドプロンプトで（XSLT を使わず）解析のみを行ってください。

Q：大量のファイルを一度に処理したら「このページのスクリプトが、Internet Explorer の実行速度を遅くしています。スクリプトの実行を続けると、コンピュータが反応しなくなる可能性があります。スクリプトを中断しますか？」という警告が出た。

---

* http://www.microsoft.com/downloads/details.aspx?familyid=2FB55371-C94E-4373-B0E9-DB4816552E41



Ａ：「いいえ」を押すことでそのまま処理を続行できます。概ね 256 ファイル以上を一度に処理する場合にこの警告が出ます。

Ｑ：短いファイルを解析すると文字化けします。
Ａ：あまりに短いファイルだと文字コードの判別に失敗することがあります。余分に文を入力するなどして、長めにして解析してみてください。

Ｑ：茶まめを強制終了したい。
Ａ：万一応答がなくなってしまった場合は、タスクマネージャを起動して mshta.exe と解析器（mecab.exe）のプロセスを終了させてください。

Ｑ：解析結果を表示するテキストエディタを指定したい。
Ａ：Internet Explorer でソースの表示をするエディタを変更してください。
（参考）@IT：Internet Explorer の［ソースの表示］メニューで起動するエディタを指定する
http://www.atmarkit.co.jp/fwin2k/win2ktips/286iesourceview/iesourceview.html