合成語生成プログラム
Word Compounder

◆ 語結び ◆

2021 年 3 月 30 日

相良かおる（西南女学院大学）

内容

1. はじめに

合成語生成プログラム「語結び」は、医療記録文に含まれる医用専門用語（合成語）を抽出し、抽出した合成語を反映した分かち書きデータを出力します。

「語結び」は Windows10（64 ビット）上で形態素解析器 MeCab とシステム辞書には IPA 辞書、そしてユーザ辞書（UTF 版の ComeJisyo）による解析結果の品詞を基に合成語を生成します。中間ファイルとして出力される MeCab による解析結果の品詞誤りをエディタなどで修正することで、利用目的に合った合成語の抽出が可能となり、利用者独自の分かち書き用ユーザ辞書の作成が容易になります。

また、生成された合成語が反映された、分かち書きデータは、機械学習用のコーパスの作成に利用することができます。図 1 は、「語結び」の利用イメージです。

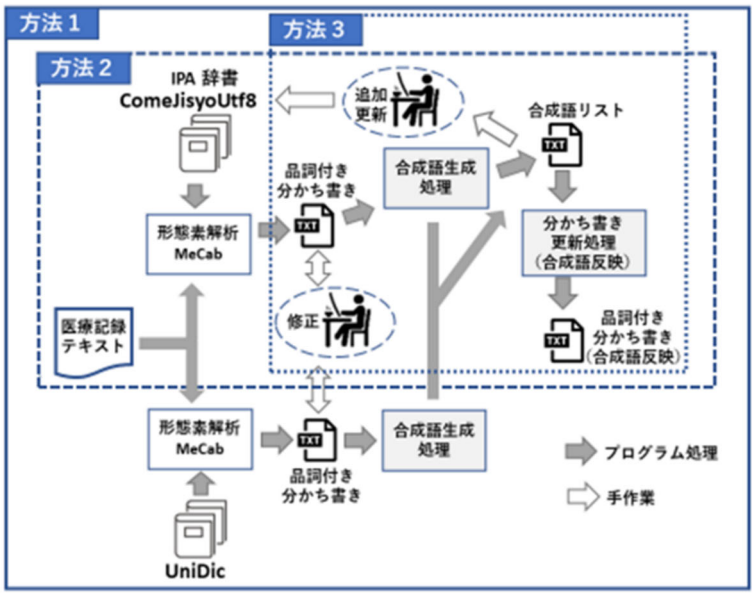

図 1　「語結び」の利用イメージ

2. 事前準備

「語結び」は Windows 10 (64bit)で動作します。本プログラムを動作させるためにはあらかじめ、形態素解析器 Mecab（UTF-8 版）をインストールし、システム辞書 IPA 辞書と、ユーザ辞書（UTF 版 ComeJisyo）を任意のフォルダーに設置してください。MeCab の Windows インストール用.exe ファイルは、以下のページの「Binary package for MS-Windows」から取得できます。

MeCab：　https://taku910.github.io/mecab/#download

ComeJisyoUtf8-2r1：　https://ja.osdn.net/projects/comedic/releases/

「語結び」は SJIS 版では動作しませんのでご注意ください。

また、本プログラムは国立国語研究所の規定した言語単位（短単位）を見出し語とした形態素解析器 MeCab 用の解析用辞書 UniDic を併用することも可能です。その場合は、任意のフォルダーに現代書き言葉 UniDic を設置して下さい。

なお、2021 年 3 月時点の unidic-cwj2.3.0.zip(2.1 GB)は、ファイルサイズが非常に大きいため、旧バージョン unidic-cwj-2.2.0.zip（439 MB）をダウンロードしてお使い頂いても構いません。

現代書き言葉 UniDic：　https://unidic.ninjal.ac.jp/download#unidic_bccwj

旧バージョン：　https://unidic.ninjal.ac.jp/back_number#unidic_cwj

3. 使用方法

(1) 「実行プログラム」のフォルダー内にある語結び WordCompounder.exe をダブルクリックして起動します（図 3）。

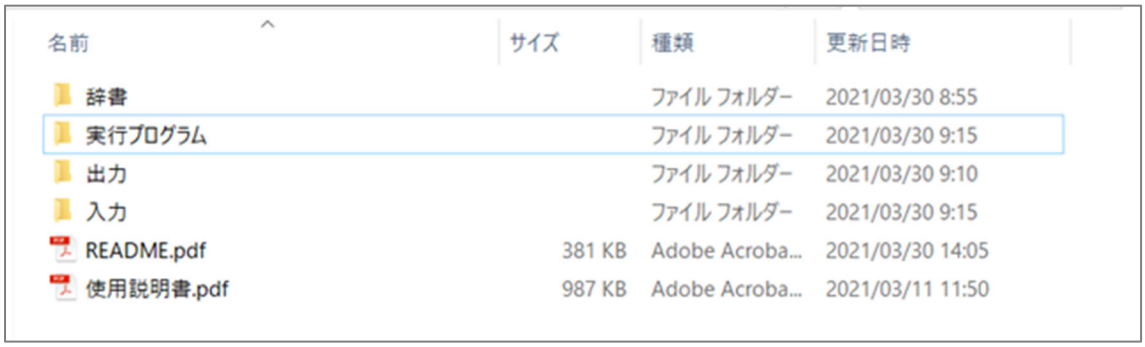

図 2　GoMusubi フォルダー

図 3　実行プログラムフォルダー

図 4. プログラムのアイコン

(2) 一般的なノート PC ですと 10 秒程待ちますと以下の黒いプロンプト画面と入力画面が表示されます。なお、黒い画面が邪魔な場合は最小化して下さい。

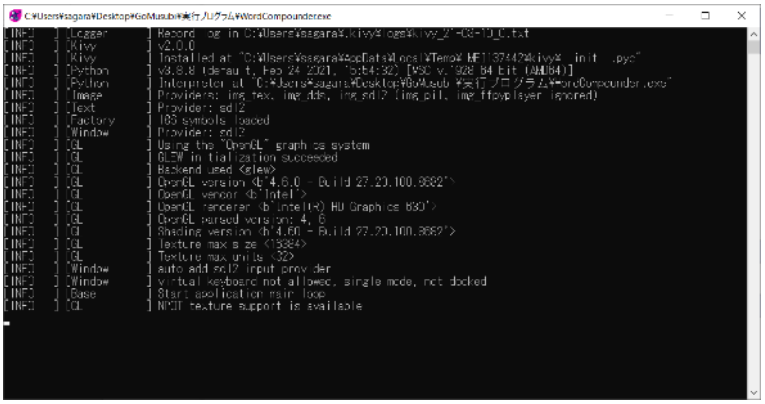

図 5. メイン画面

(3) 「参照」ボタンを押して、解析するテキストファイルを指定して下さい。次に解析結果を出力するフォルダーを指定して下さい。最後に「使用する辞書と出力先の設定」ボタンを押してください。

図 6　入出力先の設定画面

(4) システム辞書（IPA 辞書）が格納されているフォルダーとユーザ辞書のファイル名を指定して、「Close」ボタンを押してください。

出力先フォルダーや出力ファイル名を変更したい場合は指定して下さ

図 7　辞書および出力先設定画面

(5) 入力画面の「実行ボタン」を押してください。

図 8　解析成功のメッセージ（5 秒間表示）

(6) 解析に成功しますと、指定した出力フォルダーにテキスト形式の生成した合成語候補のファイルと、tab 区切りの IPA 辞書＆ComeJisyo による解析結果、そして生成した合成語による分かち書き結果の 3 種類のファイルが出力されます。

生成された合成語が反映された分かち書きデータでは、生成された合成語の品詞は「名詞,一般,合成語」となっています。

図 9 「出力」フォルダーに出力された結果データ

短単位辞書 UniDic との併用

図 10 の画面で「ComeJisyoUtf 版と UniDic」の UniDic チェックボタンをオンにし、図 11 の画面で辞書の格納場所を指定することで ComeJisyo に加えて国立国語研究所の短単位辞書 UniDic を併用することができます。短単位辞書を併用した場合、ComeJisyo と UniDic のそれぞれの解析結果から生成した合成語の中から共通する合成語を用いて分かち書きを行います。UniDic 版の実践医療用語のユーザ辞書があれば、UniDic との併用により合成語生成の精度向上が期待できます。

なお、IPA 辞書を使わず、UniDic のみでの合成語生成は出来ません。

図 10　ComeJisyoUtf 版と UniDic の併用

図 11　UniDic の設定画面

ComeJisyo 以外の辞書の利用

MeCab のユーザ辞書として利用可能な形式に変換された辞書であれば、図 7 の⑤で指定することで ComeJisyo に代わって利用することが可能です。

現在、ComeJisyo に代わって利用することの可能な辞書として、NAIST のソーシャル・コンピューティング研究室で公開されている「万病辞書」と「百薬辞書」があります。図 7 の画面の「IPAdic のユーザ辞書」でこれらの格納場所を指定することで、辞書の見出し語と品詞を元にした合成語の生成および分かち書きが可能になります。

万病辞書：　https://sociocom.naist.jp/manbyou-dic/

百薬辞書：　https://sociocom.naist.jp/hyakuyaku-dic/

形態素解析結果の直接入力

MeCab の形態素解析の精度は、使用する辞書の品詞誤りなどにより、100%ではありません。ComeJisyo の見出し語には、記号を含む合成語が含まれ、これらは過分割されます。また見出し語の通りに分かち書きされたとしても、利用者が期待する語単位ではない場合も少なくありません。そこで「語結び」では、形態素解析結果の見出し語や品詞などをエディタなどで加筆・修正したファイルを入力として合成語を生成します。利用者は、図 5 のメイン画面で「形態素解析結果を直接入力」のチェックボタンをオンにし、「IPAdic の形態素解析結果ファイル（tsv 形式）」で、加筆・修正済みの形態素解析データを設定し、「結果を出力するフォルダー」を指定し、実行ボタンを押すことで、新たに合成語を生成し、これらが反映された分かち書きデータを得ることができます。なお、一度解析したファイルを入力して再度合成語を生成することはできません。

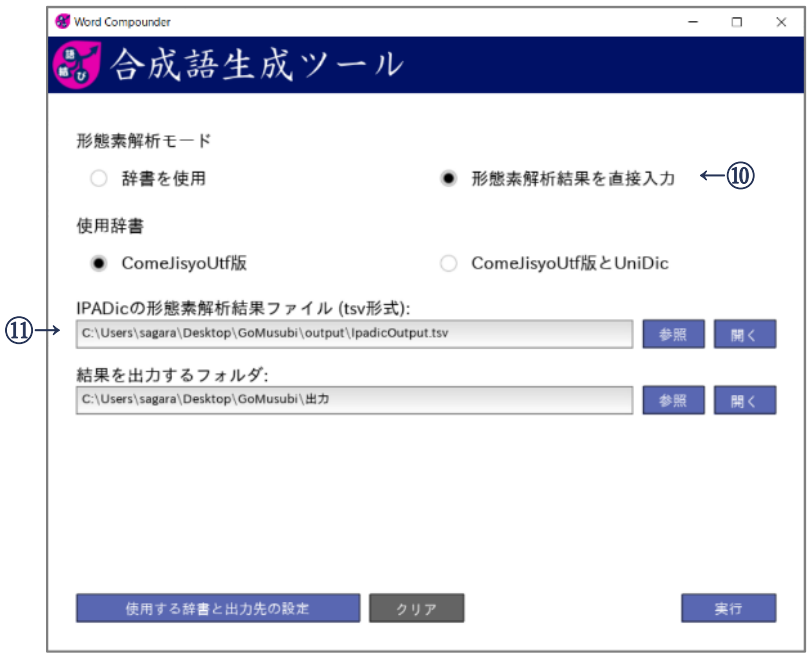

図 12　形態素解析結果からの合成語生成

辞書および出力先の設定

実行時に図 6 の画面および図 7 の画面で表示される出力先と辞書の場所は、図 3 の実行プログラムフォルダーにある設定ファイル“config.ini”の内容です。テキストエディタなどでこの内容を変更することで、画面からの入力は不要になります。

実行時のアラート表示

本ツールはオープンソースの未署名プログラムです。そのため、Windows の仕様により、初回起動時に図 13 の画像のようなアラートが表示されます。

図 13　Windows アラート画面

4. その他

不要になった場合

本プログラムが不要になった場合は、フォルダーを削除し、適宜 MeCab をアンインストールして下さい。

ライセンスと免責事項

本プログラムは、GPL/LGPL/BSD License のトリプルライセンスに従うものとします。

本プログラムを使用したことによって生じたすべての障害・損害・不具合などに関しては、一切の責任を負いません。各自の責任においてご使用ください。

なお、本プログラムは入力ファイルの大きさに制限を設けていません。本プログラムがフリーズした場合、実行画面を終了して再度実行するか、電源ボタンから再起動して下さい。

共同開発者

キュレコ株式会社　中村秋吾　https://cureco.co.jp/

株式会社システムラボラトリ　小澤泰生　https://www.algo-dev.com/

謝辞

本プログラムは、科学研究費補助金「語形成および意味的情報を付加した実践医療用語辞書の構築」（JP18H03499）の助成を受けています。