使いやすくカスタマイズ可能なテキスト解析ツールの開発

長岡技術科学大学　工学部
電気系
宮西　由貴　　山本　和英

概要

- **形態素解析器の作成を行った**

- これまでの形態素解析器の問題点
  ①形態素の区切り方が小さすぎる
  ②コマンドを打つものが多く、使い方が難しい
- 作成した解析器の特徴
  ①の解決策：使用者が好きな区切り方で区切れる
  形態素解析後に細かく区切りすぎた形態素を結合する
  ②の解決策：マウスのみで処理を行うことができる
  GUIを用いることで簡単に処理

形態素解析について

- 形態素解析の役割
  - 文を形態素に分割
  - それぞれの形態素に品詞を付与
- 形態素解析の使用例(日本語教育学)
  - 語彙リストの作成に使用

- 既存の解析器を利用する際の問題点
  - 形態素の区切りに関する問題
  - 使いやすさに関する問題

形態素の区切りに関する問題

既存の解析器における区切り方

なるべく細かく区切るようになっている

日本語教育学者にとっては、
区切るべきでない部分で区切っていると感じることも･･･[1]

例文：鈴木さんが来たんです。

欲しい結果： 鈴木 / さん / が / 来た / **んです** / 。

解析結果： 鈴木 / さん / が / 来た / **ん** / **です** / 。

改良方法は･･･

- 辞書を**作り直す**。
- 解析結果を**手動で編集**する。

大変な作業!!

[1]山内 博之:形態素解析に関する提案：日本語教育学の視点から,實践國文學 73,152-141,2008-03

使いやすさに関する問題

既存の解析器の使用方法

コマンドプロンプトで**コマンドを打つ**ことで解析される。

普段パソコンでコマンドを打たない人は、
解析器が使いにくく感じることがある。

コマンドプロンプトの欠点

マウス操作ができない
操作が視覚的に分かりにくい。
コマンドを覚える必要がある。

今回作成した解析器

目的

使いやすい形態素解析器を作成し、
多くの人に形態素解析器を使って貰う。

解析器の特徴

形態素の区切りに関する問題を
「細かく区切りすぎた形態素を**連語処理**でまとめる」
ことで解決！

使いやすさに関する問題を
「GUIを用いて**解析をマウス操作**で行えるようにする」
ことで解決！

解析の流れについて

来　:動詞
た　:助動詞
ん　:名詞
です　:助動詞
：
：
：

形態素解析部の解析結果

んです

連語リスト

連語処理部

来　:動詞
た　:助動詞
んです　:連語
：
：
：

連語処理部の解析結果

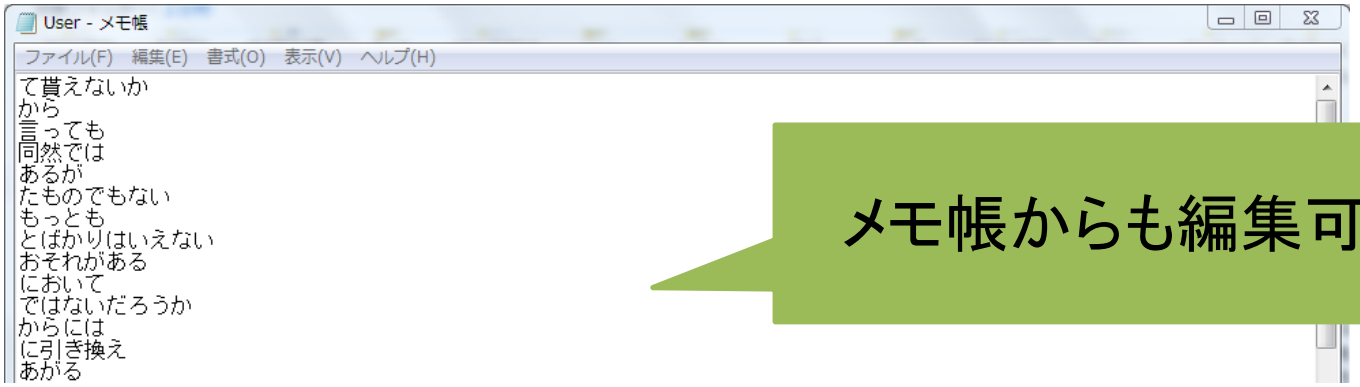
連語処理部について

- 連語処理部への入力
  - 形態素解析部の解析結果
    (現在は分かち書きの情報のみを用いている)
  - 連語リスト(後述)

- 連語処理部からの出力
  - 連語と判断された語がまとめられて出力
  - 連語と判断された語の品詞は「連語」に統一

連語リストについて

- 連語として扱いたい語を記述したリスト
- 利用者が自由に**カスタマイズ**できる



- 作成したリストの中身(見出し語のみ使用)

辞書名	説明
日本語文型辞典	日本語教育学で使用されている辞書
機能表現辞書つつじ	機能表現(2語以上で意味を成す表現)を集めた辞書

連語となる語

条件1	• 連語リスト中の語
条件2	• 含まれる形態素が2つ以上ある
条件3	• 形態素の途中から始まらない • 形態素の途中で終わらない

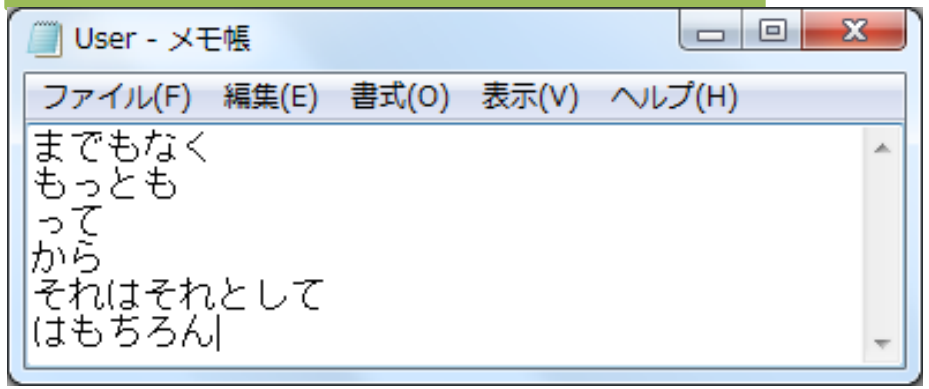
余計な表現まで取得してしまうのを防ぐ
(例) とある / がん / の / 症状 / 。
「あるが」が連語リストにあった場合に、
連語として出てしまうのを防ぐ

連語となる語

入力文(形態素解析済み)

寝坊 / を / し / て / しまっ / た / 。
今 / から / 行っ / て / も / 間に合わ / ない / こと / は / 、 / 時計 / を / 見る / まで / も / なく / 分かる / 。

連語リスト



表現	連語となるか否か	連語とならない理由
しまった	連語表現でない	連語リスト中に表現が存在しない(条件1)
から	連語表現でない	含まれる形態素が1つしかない(条件2)
って	連語表現でない	表現が形態素の途中から始まる(条件3)
までもなく	連語表現	ー

解析器の仕様

入力　テキストファイル

- メモ帳で編集が可能で、特別なエディタが必要ない

出力　CSVファイル(Excelなどでも開けるファイル)

- Excelなどを使用することで情報の抽出が容易になる

使用方法

①解析したい文をテキストファイルとして保存
②解析器にて入力ファイル＆出力ファイルを指定
③解析ボタンを押す

解析器の性能

- 形態素解析部の精度
  - 既存の解析器の結果と
    分かち書き・品詞・活用の全てが一致
- 連語処理部の精度
  - 連語リスト内の語の意味と一致

形態素解析部の精度	約97%
連語処理部の精度	約97%
解析時間(200文あたり)	約1分

まとめ

- 既存の解析器の問題点
  - 形態素の区切りに関する問題
  - 使いやすさに関する問題

- カスタマイズ可能で使いやすい解析器を作成
  - 連語処理部によって形態素をまとめた
  - GUIによって操作を簡単にした

今後の課題

- オプションの追加
  - 入出力ファイルの種類の選択
  - 出力形式を選択式 or カスタマイズ式に変更
- 連語処理部の精度向上
  - 分かち書き情報以外の情報も使用
- **ツールの公開**