セッション 4

Tregex 検索式 1

統語・意味解析コーパス (NPCMJ) チュートリアル

金城由美子

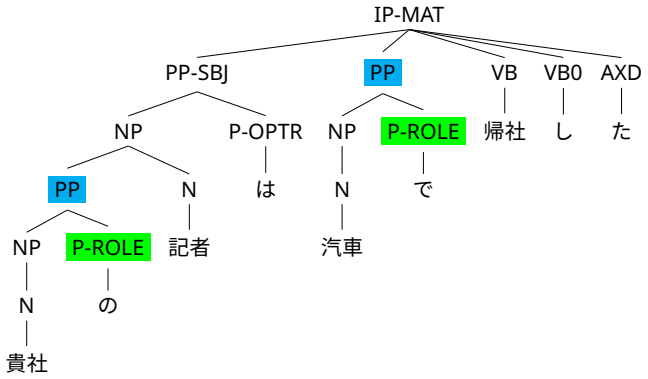
2021.3.13

セッション 4 の内容

- Tregex とは
- Tregex 検索式
- 文字列と正規表現
- 正規表現
  - 部分一致
  - 選言・グループ化
  - 任意の文字
- 検索結果のダウンロード
- ノード記述に関する補足

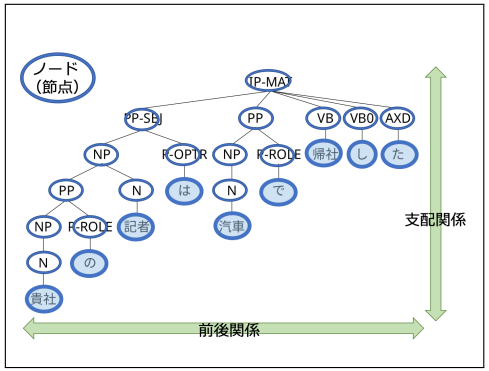
Tregex とは

- Tregex – NPCMJ のデータ（Penn 方式のデータ）を，ノードや，ノードとノードの関係を指定して検索するためのツール
- 単純な例：
  PP ＜ P-ROLE （P-ROLE を直接支配する PP）



Tregex 検索式

- 検索式は次のように分類することができる。
  - ノードを記述しただけのもの
  - 単純な関係表現：ノード記述と、それらの間の関係を 1 つ記述したもの
  - 複雑な関係表現：ノード記述と、それらの間の関係を 2 つ以上記述したもの



この時間は，ノードの記述について扱う。

文字列と正規表現

- 文字列
  単なる文字列を入力すると，
  食べ ⇒ ノード「食べ」に完全一致（「食べる」「食べ慣れる」「食べ物」などは排除される）
- 正規表現
  文字列を /.../（半角スラッシュ）で囲むと，正規表現となる。
  /食べ/ ⇒ 「食べ」に部分一致（「食べ」だけでなく，「食べる」「食べ慣れる」「食べ物」などにもマッチ）
  /SBJ/ ⇒ “SBJ” に部分一致（“PP-SBJ”，“PP-SBJ2”，“NP-SBJ”，“NP-SBJ2”，“NP;*SBJ*” などにマッチ）

練習問題 1

- 文字列検索とツリー検索で同じ文字列を検索し, 違いを確認する。ツリー検索では検索対象ファイルを指定 (”1,100p” など) すること。

正規表現

- 正規表現とは？– 文字列の集合をパターンとして表現するための記法
- スラッシュで囲むと「部分一致」の意味になる。
- 語頭（ノード先頭），語末（ノード末尾），選言，繰り返しなどを表現することができる

正規表現（部分一致）

/.../	部分一致	/食べ/ /SBJ/	「食べ」を含む “SBJ” を含む
^	ノード先頭	/^食/ /^PP/	「食」で始まる “PP” で始まる
$	ノード末尾	/する$/ /ADV$/ /^する$/	「する」で終わる “ADV” で終わる 「する」に完全一致
\b	単語の区切り	/^NP\b/	“NP”，“NP-SBJ”，“NP;{person}”，などに一致（ただし，“NPR”（固有名詞）には一致しない）

練習問題 2

- 次の検索表現はそれぞれどう違うか？(Files: 1,437p)

  (1) 首相
  (2) /首相/
  (3) /^首相/
  (4) /首相$/
  (5) /^首相$/

  (1) を確認した後，(2)〜(5) の検索結果を予想し，それぞれツリー検索画面で確かめなさい。

正規表現 (選言・グループ化)

|	選言 (A または B)	/好き|嫌い/ /SBJ|OB1/	「好き」または「嫌い」を含む “SBJ” または “OB1” を含む
(...)	グループ化	/^(好き|嫌い)$/ /^(WPRO\|WADV\|WD)$/	「好き」または「嫌い」に完全一致 “WPRO”（疑問代名詞），“WADV”（疑問副詞），“WD”（疑問限定詞）に完全一致

練習問題 3

- 動詞「思う」が使われている用例を検索するためのノードの記述の方法を考えなさい。「おもう」を考慮する。(Files: 1,100p)

  動詞の活用形の扱いについては，「データの概要とタグの検索」の補足資料，第 2 節を参照

正規表現（任意の文字 1）

.	(ピリオド) 任意の 1 文字	/^..$/ /^あ.$/	二文字の終端ノード 「あ」で始まる二文字の終端ノード
*	(アスタリスク) 直前の文字の 0 回以上の繰り返し	/^あ.*/	「あ」で始まる (“/^あ/” と同じ)
\1, \2, ... \9	検索式の中の 1〜9 番目の (...) の中身にマッチ	/^(..)\1/	始まりの二文字がもう一度繰り返される終端ノード

正規表現（任意の文字 2）

\	（逆スラッシュ）直後の文字を特殊記号ではなく通常の文字として扱う	/^*/	* で始まるもの（空要素にマッチ）
__	（アンダースコア 2 つ） ワイルドカード	__	すべてのノード ワイルドカードはスラッシュや角括弧で囲まずに使うことに注意

練習問題 4

(1) 正規表現を利用し，「〜首相」の例を探しなさい。(Files: 1,437p)

(2) 正規表現 /^ビ.*ル$/ は何を表すか？まず予想を述べ，次にツリー検索画面で確かめなさい。(Files: 100,350p)

検索結果のダウンロード

- Download all results を利用し，ダウンロードしたテキストファイルをテキストエディタ等で参照する。ツリー表示を行いたい場合は，Tregex を利用する。
- brackets 表示の画面からコピーし，エディタ・表計算ソフトにペーストする。

ノード記述に関する補足

- 正規表現では，“$” を使うか，“\b” を使うか，何も使わないかでマッチするものが変わるので注意が必要

(1a) /^NP$/
NP に完全一致。文字列 NPで検索を行うのと同じ。

(1b) /^NP/
NP で始まるすべてのノードにマッチ
（“NP”，“NP-...”，“NP;...”，“NPR”（固有名詞））

(1c) /^NP\b/
NP に完全一致するほか，NP の後に境界記号（ハイフンやセミコロン）のあるものにもマッチ
（“NP”，“NP-...”，“NP;...”）

- ハイフンは拡張タグとの境界に，セミコロンは照応情報や quantification のアノテーションに用いられる。