

主観素性を有する言語辞書を用いた ビッグデータ解析システム の研究開発

＜研究代表者＞

アーカイブ技術研究所株式会社
足立 顕

＜研究分担者＞

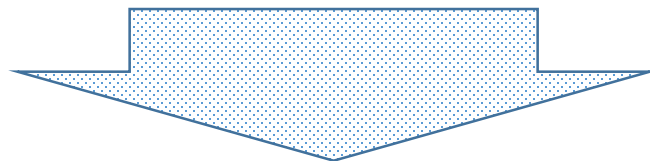
山本 竜伸
株式会社ザイナス

研究開発の内容



- …うとしない(意思)
- …かもしれない(推量)
- …にしたならば(条件)
- …になったらしい(伝聞)
- …べきなのだ(義務)

可能な付属語の連続(付属語列)を抽出



- 【研究1】 可能な付属語列の調査
- 【研究2】 付属語列は安定しているのか
- 【研究3】 付属語列辞書と形態素解析器の開発

研究開発の成果

(1) 形態素解析器を用いて抽出

【条件1】

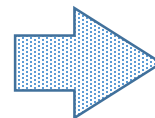
助詞、助動詞、名詞・動詞の非自立語、動詞語尾の連続

【条件2】

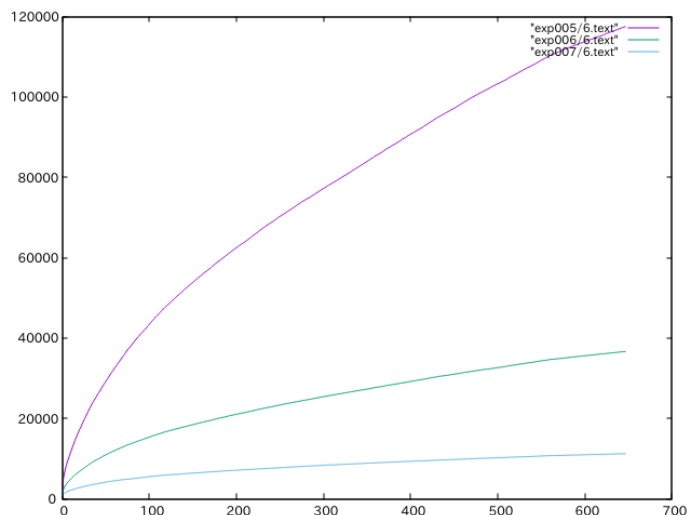
助詞、助動詞、動詞の非自立語、動詞語尾の連続

【条件3】

助詞、助動詞の連続



出現傾向の分析
人手でノイズ除去



(2) 任意の文字の連続データの作成(ツール開発)

(文字列)	(頻度)
成田空港	10
成田空港か	2
成田空港から	2
成田空港からバ	1

頻度差を利用し部分列除去

(文字列)	(頻度)
成田空港	10
成田空港から	2
成田空港からバ	1

前方から除去したものと後
方から除去しデータを作成

約430万文から約780万文字列を抽出

→ 原文に対して最長一致で適合するパターン約200万文字列のデータを作成

研究開発の成果

文内部構造解析システム (試) X 検索結果 ファイル検索システム X +

www.tech-a.co.jp/SCOPE/index.pl 80%

解析対象の日本語文を入力して送信ボタンを押してください。第一弾試作品のため最長一致したものを優先して処理している。赤字はn-gramエントリの頻度。辞書は全てのエントリ(頻度1以上)を使用。エントリの頻度は「エントリの最低頻度」の設定で変更することができます。

エントリの最低頻度: 1

夏日になるかもしれません

最長一致エントリ優先 ▼ テキストクリア 送信

↑ 処理モードで解析方式を選択できます。
最長一致エントリ優先: 解析対象文に含まれる最長のエントリから利用
先頭からの最長一致: 先頭文字から最長一致なものを選択

エントリ検索画面 ←このボタンを押すと解析用辞書を検索することができる画面を開きます
テキスト検索画面 ←このボタンを押すとKWICリストを表示する画面を開きます

夏日になるかもしれません

夏日	になるかもしれません			59
夏	日	に	なるかもしれません	5648 298815 879716 11
な	るかもしれません			570594 111
る	かもしれません			685585 382
か	もしれません			420026 385
もし	れません			6749 941
も	し	れ	ません	316330 686133 360738 18527
	ませ	ん		20049 185743
ま	せ			252661 80567

(3) データの分析ツール開発

【解析部】

- 最長一致優先選択分割
- 先頭からの最長一致分割
- 最低生起頻度指定分割

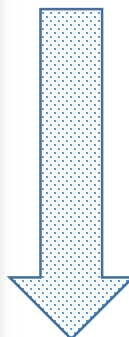
※ 文字列長のみの条件で分割

【検索部】

- エントリ検索
- KWIC検索

分割可能なところまで表示

※ 枠中、右下の数値は生起頻度



今後の取り組み

本研究開発期間では実験用システムの開発にとどまった

→ 付属語列は比較的少量の情報で高頻度のものは網羅的に取得できることが分かった

【形態素解析を用い抽出したエントリ】

→ 主観素性を付与し辞書として機能するように構築

【任意の文字の連続データ】

→ 上記エントリとの比較分析

→ 上記エントリに付与した素性を適用する枠組みの構築

【異なるタイプの言語データとの比較】

→ 今回は新聞コーパスを使用。Webデータなどとの比較。