日本語係り受けデータベース (Version 1)

変更履歴:
2010/6/7　　Version 1

1.　概要

このデータベースは、日本語の大量の Web 文書から、係り受けとその頻度を抽出したものです。ALAGIN から公開されている文脈類似語データベース Version 1.1 の生成や、参考文献[5]において本データが使用されています。

2.　ファイル

このバージョンには、データベースの本体として、

- all.sorted.nodup（全データ）
- 1m1m.sorted.nodup（全データから 100 万語に絞ったデータ）

の２つのファイルが含まれています。また、実際のファイル名は、先頭に「DEP_ALAGIN_V1_」を付加したものとなっております。

3.　ファイル容量

- all.sorted.nodup（圧縮時 4.4GB、展開後 29GB）
- 1m1m.sorted.nodup（圧縮時 1.6GB、展開後 12GB）

4.　ファイルフォーマット

それぞれのデータベースは、bzip2 で圧縮したファイルとして配布されています。（拡張子は、bz2 です）それぞれのファイルを解凍すると、元のデータベースのテキストファイルが得られます。

文字コードは UTF-8 で、正規表現で表すと以下のフォーマットで書かれていま

す。

(<語１><助詞><語 2><頻度>)¥n)+

セパレータは「半角スペース」です。

all.sorted.nodup には、795,248,234 行（約 8 億行)、
1m1m.sorted.nodup には、417,209,973 行（約 4.1 億行）が含まれています。

また、ファイルは LANG=C 環境において、UNIX コマンドの sort にてソートしたものとなっています。基本的には、<語１>が共通の行が連続することになります。

5.　生成手法

大量の Web 文書（約 1 億ページ、６０億文）を Juman/KNP により係り受け解析したデータ[3]から生成しております。この Web データは 2007 年 4 月～6 月にかけてクロールされたものです。

このデータベースは、上記の Web データから、

<語１><格助詞><語２>（野球 を 観戦する、野球 の ボール、など）

という係り受け関係を抽出し、その頻度を集計したものです。

<語１>、<語２>は文節中の複数の語の連続の場合もあります。<格助詞>についても、複数の助詞からなる場合があり、その場合には、複数の助詞を「:」でつなげたものとなっています（例えば、「から:の」など)。

基本的には、<語 1>は名詞であり、<語２>は、動詞、あるいは、助詞「の」を介して係る名詞となっています。

1m1m.sorted.nodup は、<語１>と(<助詞>,<語 2>)の種類をそれぞれ 100 万語に制限したデータで、文脈類似語データベース Version 1.1 の生成や参考文献[5]において使用されたデータと同一のものです。<語 1>については、<語 1>が現れている組(<語１><助詞><語 2><頻度>)の種類の多い上位 100 万を選択します。(<助詞>,<語 2>)については、(<助詞>,<語 2>)が現れている組(<語１><助詞><語 2><頻度>)の種類の多い上位 100 万を選択します。つまり、<語 1>としてはより多くの種類の動詞と係り受け関係にあるような名詞が選ばれることになります。

以上のように選択した <語 1>と (<助詞>,<語 2>)が両方ふくまれるような組(<語１><助詞><語 2><頻度>)のみを残したものが、1m1m.sorted.nodup です。

6. 利用条件

本データベースの利用には、(独) 情報通信研究機構と利用許諾契約を結ぶ必要があります。
詳しくは、

http://www.alagin.jp

をご覧ください。

7. 利用に関する注意

本データベースは、インターネットホームページ等、(独) 情報通信研究機構以外の第三者が作成した文書等のデータから、語彙の抽出及び統計処理等によって作成されたものです。そのため本データベースの内容は、(独) 情報通信研究機構の主体的な意思決定・判断を示すものではありません。本データベースの生成は、電子的な方法又は一様の選別基準による機械的判定によって行われています。そのため本データベースの内容の正確性、真実性及び相当性は一切保証されません。以上の理由により、(独) 情報通信研究機構は、本データベース

の内容について、責任を負いかねます。本データベースの使用に関連して生ずる損失、損害等についても、一切責任を負いかねます。本データベースには、意図せず、第三者への誹謗中傷、差別用語、個人情報などが含まれている場合があります。本データベースを利用の際はこれらによる権利侵害に十分な注意をお願いいたします。利用者においては、本データベースの以上の特質をよくご理解の上で、本データベースをご利用下さい。

8. 参考文献

[1] Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations. Jun'ichi Kazama and Kentaro Torisawa. In Proceedings of ACL-08: HLT, full poster paper, pp.407-415, June, 2008, Columbus, Ohio, USA.

[2] 係り受けの確率的クラスタリングを用いた大規模類似語リストの作成
風間淳一, De Saeger, Stijn, 鳥澤健太郎, 村田真樹 言語処理学会第 15 回年次大会 2009 年 3 月 鳥取

[3] Tsubaki: An open search engine infrastructure for developing new information access. Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. In IJCNLP 2008, 2008.

[4] ウェブ検索ディレクトリの自動構築とその改良 ---鳥式改---
鳥澤健太郎, 隅田飛鳥, 野口大輔, 柿澤康範, 風間淳一, Stijn De Saeger, 村田真樹, 黒田航, 山田一郎, 塚脇幸代, 太田公子 言語処理学会第 15 回年次大会 2009 年 3 月 鳥取

[5] "A Bayesian Method for Robust Estimation of Distributional Similarities", Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, Kentaro Torisawa, to appear in ACL 2010.

* 本データベースに関する問い合わせ先

独立行政法人情報通信研究機構
知識創成コミュニケーション研究センター
MASTAR プロジェクト 言語基盤グループ

Email: alagin-lr@khn.nict.go.jp