日本語係り受けデータベース (Version 1.1)

変更履歴:

2010/6/7　　Version 1 公開

2011/12/1　　Version 1.1 公開

1. 概要

本データベースは、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。Version 1.1 では、元となるデータを約 1 億ページ (60 億文) から約 6 億ページ (約 430 億文、クロール時期は 2007 年 5 月 19 日から 11 月 13 日) に増強し、また、Wikipedia のエントリ (記事タイトル) により形態素解析器の辞書を拡張して解析したデータを追加しました (ただし、Wikipedia エントリを含む行に関しては、本データベースとは別に http://alaginrc.nict.go.jp/wikidep/ から Wikipedia と同様のクリエイティブ・コモンズライセンスにて公開しています)。

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、

- DEP_ALAGIN_V1.1
- DEP_ALAGIN_V1.1.wikipedia.nowikipedia

の２つのファイルが含まれています。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。ファイルの容量は、それぞれ以下の通りです。

- DEP_ALAGIN_V1.1 (圧縮時 25 GB、展開後 173 GB)
- DEP_ALAGIN_V1.1.wikipedia.nowikipedia (圧縮時 25 GB、展開後 172 GB)

3. ファイルフォーマット

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。

(<表現１><助詞><表現 2><頻度>)¥n)+

データは、<表現１>が<表現２>と<助詞>を介して係り受け関係を持っている事を<頻度>回観測した事を示しています。セパレータは「半角スペース」です。

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。<表現１>が共通する行がファイル上で連続することになります。

DEP_ALAGIN_V1.1 には、4,609,525,187 行（約 46 億行）の係り受けデータが含まれています。ユニークな<表現 1>の数は 484,838,417、ユニークな<表現 2>の数は 338,355,653 です。

DEP_ALAGIN_V1.1.wikipedia.nowikipedia には、4,564,335,512 行（約 45 億行）の係り受けデータが含まれています。ユニークな<表現１>の数は 475,218,713、ユニークな<表現 2>の数は 329,163,343 です。

4.　DEP_ALAGIN_V1.1 の生成手法

大量の Web 文書（約 6 億ページ、約 430 億文）を Juman/KNP（*1）により係り受け解析したデータ（[3]の論文で述べられているデータの増強版）から生成しております。
(*1) Juman/KNP は、TSUBAKI [1] 用の内部バージョン

このデータベースは、上記の Web データの Juman/KNP の解析結果から、

<表現１><助詞><表現２>

という係り受け関係を抽出し、ノイズとなる記号・語の除去、文節の最後の動詞の原形化などの正規化処理を行った上で、各係り受けの頻度を集計したものです。なお、〈表現１〉、〈表現２〉は１文節中の複数の語の連続となることもあります

例）
野球〈を〉観戦する 40
野球〈の〉ボール 20

〈助詞〉部分は、助詞を〈〉で囲んだものになっています。係り受けが複数の助詞からなる場合には、それらの助詞を「:」でつなげたものとなっています（例えば、〈から:の〉）。また、助詞を介さない係り受けの場合には、「〈〉」として、助詞がないことを示します。

Version 1.0 とは異なり、係り元、係り先を品詞等では制限していません。そのため、より生の係り受けデータに近いデータとなっています。

5. DEP_ALAGIN_V1.1.wikipedia.nowikipedia の生成方法

係り受けデータの生成では、係り元、係り先は１文節から取り出しているため、Version 1 のデータや、上記の DEP_ALAGIN_V1.1 のデータでは「三保の松原」「風と共に去りぬ」などの２文節以上に解析されるような複雑な固有表現等に関する係り受けは含まれていませんでした。このような表現は、実際のアプリケーションで重要となるため、以下のようにして、これら語に関する係り受けデータを生成することを試みました。

- 2011 年 1 月 29 日付けの日本語版 Wikipedia のダンプデータから、Juman で解析した場合に２文節以上と解析されるエントリ（記事タイトル）を抽出（アルファベットを全て大文字化したエントリも追加）
- 上記の２文節以上のエントリの内、実際に Web 6 億ページデータに出現するエントリを抽出し、Juman の辞書に固有名詞として追加（109,745 エントリ）
- 辞書を追加した Juman と、KNP を用いて、DEP_ALAGIN_V1.1 と同じ方法で Web

6 億ページから係り受けデータを抽出

なお、Wikipedia は「クリエイティブ・コモンズ 表示-継承ライセンス」の下で公開されているため、上記のようにして抽出した係り受けデータに関しても、上記で辞書に追加した Wikipedia のエントリを係り元、あるいは、係り先に含む係り受けに関しては、別ファイルとして取り出し、Wikipedia と同様の「クリエイティブ・コモンズ 表示-継承ライセンス」の下で公開することとしました。このデータは、下記の URL から取得することができます。

http://alaginrc.nict.go.jp/wikidep/

DEP_ALAGIN_V1.1.wikipedia.nowikipedia は、辞書に追加した Wikipedia のエントリを係り元にも、係り先にも含まない係り受けデータとなります。上記の URL からダウンロードできるファイル（DEP_WIKIPEDIA_1.0）とマージすることにより、完全なデータを生成することができます。なお、辞書エントリの追加により、追加された Wikipedia エントリでない語に関する係り受けデータも DEP_ALAGIN_V1.1 とは差異が生じていますので、ご注意下さい。

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] 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.

* 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
ユニバーサルコミュニケーション研究所
情報分析研究室

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