文脈類似語データベース（Version 1.1.2）

変更履歴:

2010/6/7　　Version 1.1 （元データのバグ修正）
2010/12/10　　Version 1.1.1（各種データの追加。詳細は下記参照）
2011/1/26　　Version 1.1.2（非単語除去データの追加。詳細は下記参照）

1. 概要

文脈類似語データベースは、約 100 万語の名詞に対して、Web 文書上での文脈が類似している名詞を類似度とともに順に最大 500 個列挙したものです。

例えば、「ルパン三世」の文脈類似語の上位（括弧内は類似度）は、

ルパン３世（0.229）名探偵コナン（0.259）宇宙戦艦ヤマト（0.265）ケロロ軍曹（0.28）鉄腕アトム（0.282）ガッチャマン（0.287）デビルマン（0.289）サイボーグ００９（0.294）新世紀エヴァンゲリオン（0.295）ヤッターマン（0.305）聖闘士星矢（0.308）セーラームーン（0.308）...

のようになっていて、アニメタイトルが上位に集まっています。

また、「チャイコフスキー」の文脈類似語の上位は、

ブラームス（0.152）シューマン（0.163）メンデルスゾーン（0.166）ショスタコーヴィチ（0.178）シベリウス（0.18） ハイドン（0.181）ヘンデル（0.181）ラヴェル（0.182）シューベルト（0.187）ベートーヴェン（0.19）ドヴォルザーク（0.192）ラフマニノフ（0.193）バルトーク（0.198） ....

となっていて、有名作曲家が上位に集まっています。

一方、「カラヤン」の文脈類似語の上位は、

クレンペラー（0.21）バーンスタイン（0.215）トスカニーニ（0.227）フルトヴェングラー（0.227）ベーム（0.23）チェリビダッケ（0.232）アバド（0.239）ムラヴィンスキー（0.242）クーベリック（0.245）ヴァント（0.254）リヒテル（0.256）メンゲルベルク（0.256）ハイティンク（0.265）アーノンクール（0.276）

となっていて、有名指揮者が上位に集まっています。

（ただし、これらは例であって、全てのデータがこのように人間に理解可能なリストになっているわけではありません。詳しくは、本ドキュメント「利用に関する注意」もご参照下さい。）

Version 1.1.1 での変更点

Version 1.1.1 は、 Version 1, Version 1.1 に対する、追加データです。文献[5]で提案した新手法による文脈類似語データ、EM クラスタリング結果のモデルファイル、各語に対するクラス割り当てリスト、が追加されています。
追加ファイルだけが含まれておりますので、Version 1 および Version 1.1 のファイル、説明書もあわせてご参照ください。

Version 1.1.2 での変更点

これまで配布してきた文脈類似語データベースは、自動解析された結果から単語を抽出しているため、実際には意味をなさない文字列（非単語）が単語として認識されていることがありました。そこで、約 100 万の単語候補から人手で確認した約 33,000 の非単語を削除したデータを作成し、本バージョンで配布いたします。ただし、全ての非単語を除去した保証はなく、残った単語に非単語が存在することがあります。また、正しい単語を非単語として削除している場合もございますので、ご了承下さい。追加ファイルだけが含まれておりますので、Version 1 および Version 1.1, 1.1.1 のファイル、説明書もあわせてご参照ください。

2. ファイル

このバージョンには、以下のデータが含まれています。

- **SW_ALAGIN_V1.1.1_1m-rv100k.bbc{0.0008,0.0016}.cleaned.data.bz2**
  - Version 1.1.1 で配布された SW_ALAGIN_V1.1.1_1m-rv100k.bbc{0.0008,0.0016}.data.bz2 から、非単語を取り除いたデータです。（以下、1m-rv100k.bbc{0.0008,0.0016}.cleaned.data と略記）
- **SW_ALAGIN_V1.1_1m-2k.s1+s2.cleaned.data.bz2**

- Version 1.1 で配布された、SW_ALAGIN_V1.1_1m-2k.s1+s2.data.bz2 から、非単語を取り除いたデータです。(以下、1m-2k.s1+s2.cleaned.data と略記)
-

- SW_ALAGIN_V1.1_1m-rv100k.cleaned.data.bz2
  - Version 1.1 で配布された、SW_ALAGIN_V1.1_1m-rv100k.data から非単語を取り除いたデータです。(以下、1m-rv100k.cleaned.data と略記)

3. ファイル容量

- 1m-rv100k.bbc0.0008.cleaned.data (圧縮時 2.4GB、展開後 9.4GB)
- 1m-rv100k.bbc0.0016.cleaned.data (圧縮時 2.3GB、展開後 9.0GB)
- 1m-2k.s1+s2.cleaned.data (圧縮時 2.6GB、展開後 11GB)
- 1m-rv100k.cleaned.data (圧縮時 2.6GB、展開後 11GB)

したがって、全てのファイルの格納には圧縮時で約 9.9GB 展開後で約 40.4GB が必要となります。

4. Version 1.1.2 のデータ詳細

拡張子が bz2 の場合は、bzip2 で圧縮したファイルとして配布されています。その場合には、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。

文字コードはすべて UTF-8 で、ファイルフォーマットは、以下の正規表現で表すことができます。

(<対象名詞>¥n(<類似名詞> <類似度>)+¥n)+

１つの名詞に対する記述は２行にわたっており、最初の行にその名詞、次の行に文脈類似語の情報が続きます。

例えば、

りんご
みかん 0.9 バナナ 0.5 パイナップル 0.2 ...
犬
猫 0.7 たぬき 0.6 猿 0.5 ...
...

のような形式をしています。

上の例では、「りんご」に最も文脈が類似している名詞は「みかん」であり、類似度は 0.9、「犬」に一番類似している名詞は「猫」であり、類似度は 0.7、というように読み取ることができます。詳しくは、Version 1, 1.1, 1.1.1 のマニュアルを参照ください。

4.1. 非単語削除の方法

元のファイルの<対象名詞>が非単語の場合には、その対象名詞に関するデータすべて（2 行分）が削除されています。<類似名詞>が非単語の場合には、その非単語と類似度がデータから削除されています。<対象名詞>-<類似名詞>の組として捉えた場合の削除数は以下の通りです。全データ量の 1～2%がこの処理により取り除かれています。

データ名	削除組数/元の組数
1m-rv100k.bbc0.0008.cleaned.data	5,786,454 / 498,360,649
1m-rv100k.bbc0.0016.cleaned.data	5,094,581 / 498,360,843
1m-2k.s1+s2.cleaned.data	8,982,792 / 498,867,088
1m-rv100k.cleaned.data	9,501,463 / 498,863,296

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。

5. 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。

6. 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
上記論文で述べられている手法の一部のよりオリジナルの論文としては、以下の論文があります。
PLSI Utilization for Automatic Thesaurus Construction, Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama, IJCNLP 2005.

[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, ACL 2010.

[6] Large Scale Relation Acquisition using Class Dependent Patterns, Stijn De Saeger, Kentaro Torisawa, Jun’ichi Kazama, Kow Kuroda and Masaki Murata, In Proceedings of the IEEE International Conference on Data Mining (ICDM’09), pp.764-769, December, 2009, , Miami, Florida, USA.

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