文脈類似語データベース（Version 1）

変更履歴:
2009/12/17　（一括ダウンロードファイルの説明を追加）
2010/4/2　（ファイル容量について追加）

＊ 概要

このデータベースは、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）...

となっていて、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、これらは一例であって、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。）

＊ 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、

- 1m-2k.s1.data（100 万語, 2000 クラスのクラスタリングモデルによるリスト）
- 1m-2k.s2.data（上記の、クラスタリング時の初期値違い）
- 1m-2k.s1+s2.data（1m-2k.s1 のモデルと 1m-2k.s2 のモデルを組み合わせて生成したリスト）
- 1m-rv100k.data（100 万語, 動詞 10 万種類との共起に基づいて生成したリスト）
- old.500k-2k.data（50 万語, クラスタリングモデル）

の 5 つのデータベースが含まれています。
(配布の際には、ファイル名に全て、SW_ALAGIN_V1_ という文字列が付与されています。例えば、SW_ALAGIN_V1_1m-2k.s1.data_01.bz2 のようなファイル名になっています。説明を煩雑にしないため、下の説明では、それを省略していますが、適宜補ってご理解ください)

それぞれ、文脈類似語を列挙する方法が異なっています。詳しくは、「生成手法」の項を参照してください。

＊ ファイル容量
- 1m-2k.s1.data（圧縮時約 2.7GB、展開後約 11GB）
- 1m-2k.s2.data（同上）
- 1m-2k.s1+s2.data（同上）
- 1m-rv100k.data（同上）
- old.500k-2k.data（圧縮時約 1.3GB、展開後約 4.3GB）

したがって、全てのファイルの格納には圧縮時で約 12.1GB 展開後で約 48.3GB

が必要となります。

* ファイルフォーマット

それぞれのデータベースは、一括してダウンロードできる bzip2 で圧縮したファイルと、大きなファイルをダウンロードできない場合のために、20 万行毎（10 万語分、ただし最後のファイルは端数あり）に分割した上で、bzip2 で圧縮したファイルの二種類があります。

例えば、1m-2k.s1.data は一括の場合は、

1m-2k.s1.data.bz2

分割の場合には、

1m-2k.s1.data_01.bz2
1m-2k.s1.data_02.bz2
...
というファイルに分割されています。

それぞれのファイルを解凍し、UNIX の cat コマンドなどでファイル名の数字の順番でつなげると、元のデータベースのテキストファイルが得られます。(つなげなくても別々で使うこともできます)

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。

(<名詞>¥n(<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 -0.5 バナナ -0.8 パイナップル -0.9 ...
犬
猫 -0.1 たぬき -0.2 猿 -0.7 ...
...

のような形式をしています。

上の例の場合、「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-0.5、「犬」に一番類似している名詞は「猫」であり、類似度は-0.1、というように読み取ることができます。

類似度は、次の生成手法の項で述べるように、
Jensen-Shannon divergence にマイナスを付けたものになっています。また、文脈類似語は、類似度の大きい順に書かれています。

* 生成手法

** 元データの生成

大量の Web 文書を係り受け解析したデータ[3]から、

名詞 1 助詞 動詞 （「野球 を 観戦する」など）
名詞 1 助詞 名詞 2（「野球 の ボール」など）

という「名詞と動詞」、「名詞と名詞」という２つのタイプの係り受け関係を抽出し、それを「名詞 1」の部分の名詞に対する文脈として用います。

これらの係り受けは、Web 文書全体で集計されて、

(名詞 1 動詞@助詞 頻度)
(名詞 1 名詞 2@助詞 頻度)

という三つ組みにそれぞれ変換されます。ただし、プログラム上は、
n = 名詞 1, vt = 動詞@助詞 または 名詞 2@助詞, f(n,vt)を Web 中での頻度と定義して、(n vt f(n,vt)) という形式のデータとして区別なく扱っています。

このデータから、

n については、n が現れている (n vt f(n,vt)) の種類の多い上位 100 万を選択します。vt については、vt が現れてる (n vt f(n vt)) の種類の多い上位 100 万を選択します。(つまり、n 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 n と vt を両方に含むような (n vt f(n,vt)) のみを、次に述べるデータベースの生成の際に使用します。

したがって、データベースに含まれている名詞の数は、100 万よりわずかに小さくなっており、実際には、999,757 です。また、データベースでは、上記の選択の順番で対象の名詞が出現します。

** 1m-2k.s1.data、1m-2k.s2.data の生成方法

上の係り受けデータを、論文[1]で述べられている手法でクラスタリング(クラス数=2,000)し、論文[2]の「手法 A」で、M=1,600 N=500 として生成したものです。

したがって、各々の名詞には最大 500 個の文脈類似語が列挙されています。(ただし、出力が 500 個に満たない場合もあります)

ここで用いている係り受けデータは、論文[2]で用いられていたデータとは名詞・動詞抽出基準の変更、バグの修正、語として不適切な文字列を取り除く処理の追加のために、多少異っています。また、細かいことですが、クラスタリングの確率値の閾値パラメータが、10E-6 から 10E-10 に変わっています。

論文[2]の「手法 A」では、クラスタリングの結果得られるクラス所属確率分布 p(c|n) の間の Jensen-Shannon divergence によって名詞間の距離を計算し、マイナスをかけたものを類似度としています。具体的な式は、

sim(n1, n2) = - JS(n1 || n2)

JS(n1 || n2) = 0.5 * KL(p(c|n1) || (p(c|n1) + p(c|n2))/2)
+ 0.5 * KL(p(c|n2) || (p(c|n1) + p(c|n2))/2)

KL(p(c|n1) || p(c|n2)) = ¥sum_c p(c|n1) log (p(c|n1)/p(c|n2))

となります。以上で計算される類似度は、0 から-1.0 の範囲の数値となります。

クラスタリングの初期値の生成には、乱数がつかわれており、s1 は乱数の種を 1 にして生成した場合、s2 は 2 にして生成したものです。初期値の差により、結果の文脈類似語リストも変化します。

** 1m-2k.s1+s2.data の生成方法

これは、論文[2]の「手法 C」を用いて、1m-2k.s1 のクラスタリングモデルと 1m-2k.s2 のモデルを組み合わせて同時に使用してリストを生成したものです。

論文[2]の人手評価による実験では、二つのモデルの平均をとることにより、より精度の高いリストが得られることが観察されています。

** 1m-rv100k.data　について

こちらは、上で述べたようなクラスタリングは行わずに、元データを直接用いて生成したものです。

p(n|vt) = log f(n,vt) / ¥sum_n log f(n,vt)　（式 1)
p(vt)　= ¥sum_n log f(n,vt) / ¥sum_vt ¥sum_n log f(n,vt) (式 2)

という式を用いてこれらの確率値を推定し、ここから、p(vt|n)をベイズ則を用いて求め、それがあたかも上の「手法 A」での p(c|n)であるかのようにして、類似度を計算したものです（上と同じく、M=1600 N=500)。

（ただし、メモリ量と計算量を現実的な規模に抑えるため、vt は上位の 10 万に制限して計算しています。)

（式１)、(式 2)　で頻度ではなく「頻度の log」を用いているのは、文脈類似語の精度を向上させるためです。この手法は、1m-2k.s1 や 1m-2k.s2 のクラスタリングの際にも用いられています。効果の詳細については論文[2]を参照ください。

1m-2k.s1, s2, s1+s2 と、1m-rv100k の違い:

クラスタリングを用いない 1m-rv100k の方が、詳細で文脈が的確に類似している語が上位になる傾向があるようです。ただし、用途によっては、より幅広い語が上位になっている 1m-2k.s1+s2 のほうが適しているかもしれません。用途により使い分けていただければと思います。

** old.500k-2k.data について

生成手法は、1m-2k.s1,s2,s1+s2 などと同等ですが、語彙数は 50 万語で、M=1000 N=500 というパラメータで生成してあります。他のデータより古いバージョンの文脈類似語のデータベースです。係り受け抽出部分のバグなどが修正されていない時点でのデータ（例えば、英語などの語間のスペースが削除されてしまっている）ですが、これまで行った各種展示会や発表などではこのデータに基づいて説明を行っていた場合もありますので、参考のため、合わせて公開をいたします。

* 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。

* 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。

* 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

* 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