文脈類似語データベース（Version 1.1.1）

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）

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 から Version 1.1.1 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。

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}.data.bz2**
  [1m-rv100k.bbc{0.0008,0.0016}.data と略記]
  - 文献[5]で提案した新手法による文脈類似語リストです。それぞれ、alpha=0.0008 および、0.0016 で生成したものです。
- **SW_ALAGIN_V1.1.1_1m-{5h,2k}.{s1,s2}.model.bz2**
  [1m-{5h,2k}.{s1,s2}.model と略記]
  - 文献[2]で述べている EM アルゴリズムによる単語クラスタリング手法で生成した確率モデルファイルです。5h はクラス数が 500 であること、2k は 2,000 であることをあらわしています。s1, s2 は、初期値が違うことをあらわしています。2k の s1,s2 は、Version 1.1 に含まれていた類似語リスト「SW_ALAGIN_V1.1_1m-2k.s1+s2.data.bz2」の生成の大本になっているファイルです。文脈類似語リストの生成以外にも、様々な用途に利用することが可能なデータです。
- **SW_ALAGIN_V1.1.1_1m-{5h,2k}.{s1,s2}.class** [1m-{5h,2k}.{s1,s2}.class と略記]
  - 上記のモデルファイルを用いて、各名詞に隠れ意味クラスを割り振ったファイルです。この意味クラスを用いてアプリケーションの動作を制御するなど、いくつかの利用法が考えられます。

3. ファイル容量

- 1m-rv100k.bbc0.0008.data（圧縮時 2.5GB、展開後 9.8GB）
- 1m-rv100k.bbc0.0016.data（圧縮時 2.3GB、展開後 9.3GB）
- 1m-2k.s1.model（圧縮時 906MB、展開後 2.8GB）
- 1m-2k.s2.model（圧縮時 906MB、展開後 2.8GB）
- 1m-5h.s1.model（圧縮時 430MB、展開後 1.3GB）
- 1m-5h.s2.model（圧縮時 432MB、展開後 1.3GB）
- 1m-2k.s1.class（圧縮時 9.3MB、展開後 30MB）
- 1m-2k.s2.class（圧縮時 9.4MB、展開後 30MB）
- 1m-5h.s1.class（圧縮時 9.4MB、展開後 31MB）
- 1m-5h.s2.class（圧縮時 9.6MB、展開後 31MB）

したがって、全てのファイルの格納には圧縮時で約 7.5GB 展開後で約 27.5GB が必要となります。

4. Version 1.1.1 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 です。

4.1. 1m-rv100k.bbc{0.0008,0.0016}.data

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。

(<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 0.9 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。このデータでは、Version 1, Version 1.1 のデータとは異なり、類似度は 0 から 1 の間の値を取っています。

（注意）
生成には、文献[5]で提案したベイズ手法（BBC 法）を用いています。実験では、評価データによっては既存の手法よりも高い精度で文脈類似語を見るけることができことが分かっています。しかし、実際のアプリケーション使用する場合に、これまで配布してきた文脈類似語リストと比べて必ずしも性能が向上することを保証するものではありません。用途によって、使い分けなどをしてください。全般的な傾向としては、BBC 法によって生成した場合には、頻度が高い語が類似語として出力される傾向があります。0.0008 よりも 0.0016 のほうがその傾向は強くなります。また、プログラムの違いから、Version 1, Version 1.1 で配布した文脈類似語データベースとは含まれる語が多少異なりますので、ご注意ください。

4.2. 1m-2k.{s1,s2}.model の詳細

4.2.1. 元データの生成

まず、本データを生成するために、以下のように元データを作成しました。大量の Web 文書（約一億ページ、６０億文）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。Vt のことを、「動詞テンプレート」と呼びます。

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います。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。

なお、上記の係り受けデータおよびそのもととなった Web から抽出した係り受けデータは、ALAGIN より、「日本語係り受けデータベース Version 1」として公開されております。あわせてご利用下さい。

4.2.2.　クラスタリング

前節で説明した係り受けデータを用いて、EM アルゴリズムを用いてクラスタリングを行います。詳細は文献[2]をご覧ください。使用している確率モデルは

$$p(n, vt, c) = p(c)\,p(n \mid \mathrm{c})\,p(vt \mid \mathrm{c})$$

という式で表されます。ここで、c は、隠れクラス(単語の意味クラス)です。クラスタリングの結果、 $p(c), p(n \mid \mathrm{c}), p(vt \mid \mathrm{c})$ の確率値がパラメータとして得られます。モデルファイルにはこれらの値が格納されています。隠れクラスはモデルファイル上は単なる整数の id で表されており、意味を表す名前がつけられているわけではありませんが、文献[2]にあるように、文脈類似語データを生成したり、文献単語をクラス分けするなど、様々に利用することができます。

ファイル名の s1, s2 は、EM アルゴリズムを異なる初期値で実行したことを表します。EM アルゴリズムでは、初期値によって得られる結果が異なることが知られており、文脈類似語の生成では、これら二つのモデルでの類似度を平均することでより高い精度となることが分かっています。注意としては、s1 と s2 での隠れクラスには何の関係もないということです。つまり、s1 でのクラス 100 の意味は s2 でのクラス 100 の意味とは一致しませんので、利用の際にはご注意ください。

4.2.3.　ファイルフォーマット

クラスタリングモデルファイルは、以下のフォーマット形式となっております。

N	← 名詞数
T	← 動詞テンプレート数
C	← クラス数
n_0	← 名詞の文字列（N 行続きます）
…	0 ~ N-1 の整数が各名詞の id となります。
n_{N-1}	
vt_0	← 動詞テンプレートの文字列（T 行続きます）
...	0 ~ T-1 の整数が各名詞の id となります。
vt_{T-1}	
p(c_0)	← クラス生成確率 p(c)（C 行続きます）
...	
p(c_{C-1})	
#nz_p(n|c_0)	← c_0 に対して、確率値が閾値 H 以上の n の個数
n_i1	← 確率値が閾値 H 以上の一つ目の n の id
p(n_i1|c_0)	← 確率値。この 2 行が#nz_p(n|c_0)個だけ続きます。
さらに続けて c_1, ... c_{C-1}まで同様の記述が続きます。	
#nz_p(vt|c_0)	← c_0 に対して、確率値が閾値 H 以上の vt の個数
vt_i1	← 確率値が閾値 H 以上の一つ目の vt の id
p(vt_i1|c_0)	← 確率値。この 2 行が#nz_p(vt|c_0)個だけ続きます。
さらに続けて c_1, ... c_{C-1}まで同様の記述が続きます。	

モデルファイルの容量を削減するため、p(n|c)と p(vt|c)は、確率値が H 以上のもののみ書き出しています。配布するファイルでは、H = 1.0E-10 となっております。注意としては、以上のように、確率値を捨てていますので、モデルファイルの確率値を合計しても１にならない場合があります。そのような想定をしているプログラムで使用する場合は、確率値を足して１になるように補正（ノーマライズ；合計した値で各値を割ること）して使用してください。

このモデルファイルを使用して、例えば、各名詞に対する隠れクラスの分布p(c|n)などを以下のようにして求めることができます（ベイズの定理)。

$$p(c \mid n) = \frac{p(c)\,p(n \mid c)}{\sum_{c'} p(c')\,p(n \mid c')}$$

この隠れクラスの分布に対して、Jensen-Shannon ダイバージェンスなどの確率分布に対する距離計算を行えば、名詞の文脈類似度を計算したりできます。
また、上記確率が一番大きいクラスなど見つけることによって名詞の意味的な分類も行えます。次に説明する 1m-2k.s1.class などはまさしくそのようにして生成したものです。

4.3. 1m-2k.{s1,s2}.class の詳細

このデータは、各名詞 n に対して、p(c|n)が閾値（0.2）以上となる隠れクラスの番号と確率値を列挙したものです。閾値を越える隠れクラスがない場合は、確率値が最大の隠れクラスが出力されています。

ファイルフォーマットは以下のとおりです。

---

n_0
c_i1<SPC>p(c_i1|n_0) .... c_iK<SPC> p(c_iK|n_0)
...　上記 2 行が名詞数だけ続きます
n_{N-1}
c_i1<SPC>p(c_i1|n_{N-1}) .... c_iK'<SPC> p(c_iK'|n_{N-1})

1m-2k.s1.class は 1m-2k.s1.model から、1m-2k.s2.class は 1m-2k.s2.model から生成されています。ここでも、s1 と s2 のクラス id 間には何の関連もないのでご注意ください。このようなデータは、例えば文献[６]などで、意味的知識を獲得する際の意味的制約として利用されており、その有効性が示されています。

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