日本語パターン言い換えデータベース（Version 1）

変更履歴
2010 年 4 月 05 日: データ容量を追記
2009 年 12 月 21 日: （最初バージョン）

1. 概要

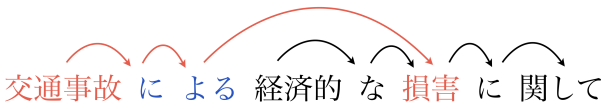
本データベースは、文の係り受け解析の結果を利用して、「A が B の原因となる」というような､文内に任意の名詞 A と B を結ぶ表現パターンの言い換えデータベース(各々のパターンに対して類似したパターンをその類似度とともに列挙したもの) です。例えば、〈A は B の原因となる〉という表現に関しては下記のものがデータベースの出力となります。

<A は B の原因になる> 0.0578512397	<A すると B の原因となる> 0.0155563071
<A は B の原因> 0.0400322407	<B の原因といわれる A> 0.0155361879
<B の原因である A> 0.0370898716	<A が B の原因である> 0.0137732223
<A は B の原因にもなる> 0.0346598203	<B の原因の A> 0.0136869119
<B の原因となる A> 0.0335473370	<B は A が原因です> 0.0133979475
<B の原因になる A> 0.0328618063	<B の原因でもある A> 0.0132726583
<B の原因にもなる A> 0.0222953216	<B の元となる A> 0.0131449872
<A が B の原因となる> 0.0220117024	<A は B の原因になるので> 0.0130890052
<B の原因とされる A> 0.0206469374	<A を原因とする B> 0.0130201685
<A が B の原因になる> 0.0204869359	<A によって起こる B> 0.0130054867
<A は B を引き起こす> 0.0203133298	<A が B の原因になること> 0.0128544423
<B の原因は A です> 0.0197772221	<B の元になる A> 0.0124058485
<A が B の原因> 0.0197204194	<A は B の元になる> 0.0123367199
<B などの原因となる A> 0.0191421482	<B は A が原因> 0.0122401848
<A が原因で起こる B> 0.0191314910	<B のもととなる A> 0.0119910067
<A が原因となる B> 0.0186186186	<A は B のもと> 0.0118072748
<A が B を引き起こす> 0.0179789213	<A は B の原因となるので> 0.0117379780
<A は B の原因です> 0.0175746924	<B の原因の一つである A> 0.0117246596
<B の原因 A> 0.0173611111	<B を引き起こす A> 0.0114091890
<B の原因となっている A> 0.0169395822	<B の原因のひとつである A> 0.0108737864
<B の原因は A> 0.0166882463	<A は B の原因ともなる> 0.0105860113
<A が引き起こす B> 0.0157584963	<AB の原因> 0.0105288346

<A すると B の原因になる> 0.0156943023	<B の主な原因は A です> 0.0104829652
<B の原因ともなる A> 0.0156366344	<A も B の原因になる> 0.0104821803
<B のもとになる A> 0.0156104380	<A は B の元> 0.0104801365
<A すると B の原因となる> 0.0155563071	<B の要因となる A> 0.0104675506
<B の原因ともなる A> 0.0156366344	<A によって引き起こされる B> 0.0103116407
<B のもとになる A> 0.0156104380	<A は B を招く> 0.0103032087

2. 作成方法

この説明書ではこれらの言語表現を「パターン」と呼びます。厳密には、パターンが係り受け解析の結果となる構文木の中で、一定の出現頻度を超える名詞 A と B をつなぐ係り受けパスに含まれる単語からなります。例えば、「交通事故による経済的な損害に関して」という文から<A による B>というパターンが抽出されます。



パターン：<AによるB>

本データベースは 5000 万ウェブ文書（[1]）から獲得したパターンを言い換えの対象とします。パターンの類似度は各パターンと共起する名詞対の統計を用いて、パターン間の Jaccard coefficient (http://en.wikipedia.org/wiki/Jaccard_index)として計算します。詳しくは[2]をご覧ください。本データベースのパターン類似度は、[2]の「SC」（“Single Class”）という比較手法で使われたパターン類似度の尺度に相当します。

なお、本データベースで「同じ名詞対と共起するパターンが意味的に類似している」という仮説に基づいてパターンの言い換えを列挙します。コーパス上の語の共起頻度を用いて自動的に作成したもので、明らかに言い換えでないパターンも含まれていますのでご了承ください。例えば、下記の実行例の中、「<A は B が嫌い>」というパターン（最後の例）はその反対の意味を持つ「<A は B が好き>」というようなパターンが言い換えとして生成されてしまいます。

[1] K. Shinzato, T. Shibata, D. Kawahara, C. Hashimoto and S. Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access. In Proceedings of IJCNLP, 2008. pp. 189-196

[2] S. De Saeger, K. Torisawa, J. Kazama, K. Kuroda, M. Murata. Large Scale Relation Acquisition using Class Dependent Patterns. Proceedings of ICDM, 2009. pp. 764

3. ファイル

本データベースは検索スクリプトとデータファイルからなります。データファイルの各行のフォーマットは次の通りです：

〈パターン１〉〈パターン２〉［コーパスでパターン１と共起する名詞対のことなり数］［コーパスでパターン２と共起する名詞対のことなり数］［コーパスでパターン１とパターン２、両方と共起する名詞対のことなり数］

例：

〈A していることを B されました〉〈A する可能性が B されています〉 40 103 2
〈A していることを B されました〉〈A する可能性が B されている〉 40 177 4
〈A していることを B されました〉〈A する可能性があることが B された〉 40 31 2
〈A していることを B されました〉〈A する可能性も B されている〉 40 83 2

データファイルの展開と連結

scripts というディレクトリに reconstruct-data.sh というスクリプトがあります。ダウンロード後にそれを用いてデータファイルを展開し、作り直します。展開前のデータ容量が約 22GB（約 4GB のファイルを５個）です。展開後のデータ容量が約 157GB（167898126408 バイト）になりますので、ディスク容量にご注意ください。

```
$ cp scripts/reconstruct-data.sh path/to/data/files/
$ cd path/to/data/files/
$ bash reconstruct-data.sh
reconstructing data...
done
checking reconstructed data...(may take a while)
done
$
```

同ディレクトリにある検索スクリプト(find-similar-patterns.pl)の方は、ユーザが入力するパターンで言い換えの候補となるパターンを検索して、各候補の類似度を計算して、類似度が高い順で出力します。スクリプトを実行するには Perl (www.perl.com) が必要です。なお、スクリプトが二分探索でデータを検索しますので、Unix 系の OS 上で system locale によってソート順が影響されることがあります。場合によって“LANG”, “LC_ALL” という環境変数を “C” に設定する必要があります (Bash なら、export

LANG=C; export LC_ALL=C)。

実行コマンド：$ perl find-similar-patterns.pl path/to/patternDB.dat arg

プログラムの引数はデータファイル（patternDB.dat）と検索したいパターンやパターンファイルです(arg)。

実行例：

```
$ perl scripts/find-similar-patterns.pl ./patternDB.dat '<A は B が豊富です>'
<A は B が豊富> 0.0549719888
<A には B が豊富に含まれています> 0.0382925298
<A は B も豊富です> 0.0377786173
<A は B を多く含む> 0.0336538462
<A は B も豊富> 0.0331325301
<B を豊富に含む A> 0.0314937013
<A には B が多く含まれています> 0.0287745430
<B が豊富な A> 0.0270556518
<A は B を豊富に含んでいます> 0.0269879518
<A は B 豊富> 0.0269503546
<A は B が豊富に含まれています> 0.0268899036
<A に豊富に含まれる B> 0.0267541646
<A は B が豊富で> 0.0258192651
<A は B を豊富に含む> 0.0255075482
<B 豊富な A です> 0.0231871838
<A は B を多く含んでいます> 0.0226551880
<A には B がたっぷり含まれています> 0.0225563910
<A に多く含まれる B> 0.0223623853
…
$ perl find-similar-patterns.pl ./patternDB.dat '<A 用の B>'
<A のための B> 0.0262483995
<A 用の B です> 0.0254558178
<B を使った A> 0.0198032617
<A 用の B とする> 0.0174796903
<AB とする> 0.0171181411
<A するための B> 0.0163295657
<A に使う B> 0.0161732743
<A に必要な B> 0.0160494507
<A できる B> 0.0141996481
```

<AB です> 0.0135189382
<B で A できる> 0.0124584114
<B を使って A する> 0.0123363896
<B の A 用> 0.0119705990
<A の B とする> 0.0116086159
<B で A をする> 0.0112237818
<A で使う B> 0.0110206379
<A に使用する B> 0.0104004238
<B を A できる> 0.0101187967

```
$ perl find-similar-patterns.pl ./patternDB.dat '<A は B を防ぐ>'
```
<A が B を防ぐ> 0.0224161276
<A は B を予防する> 0.0186121788
<A で B を防ぐ> 0.0175963197
<B を防ぐ A> 0.0175141447
<A は B を防止する> 0.0132786565
<B を予防する A> 0.0132532850
<B を防ぐ A です> 0.0118343195
<B を防止する A> 0.0117291936
<A に B を防ぐ> 0.0114255581
<A は B を防いでくれます> 0.0108145421
<AB を防ぐ> 0.0103763358
<A が B を予防する> 0.0102552913
<A は B を抑える> 0.0102269543
<A は B を防いでいます> 0.0101786926

```
$ perl find-similar-patterns.pl ./patternDB.dat '<A で B を喜ばせる>' | head -30
```
<A を B 様にご提供していきたい> 0.0430107527
<B 様に A を提供して参りました> 0.0337078652
<A を B 様に提供し続けること> 0.0337078652
<B 様に A を提供出来るように> 0.0337078652
<B 様に A を提供出来るよう> 0.0333333333
<B 様に A を約束する> 0.0309278351
<B 様に A をお楽しみいただける> 0.0309278351
<B 様に A をご提供したい> 0.0300000000
<B 様に A をご提供できる> 0.0298507463

<A を味わいたい B 様> 0.0297029703

<A を B 様にご提供できるように> 0.0297029703

<B 様一人一人のニーズに合わせた A> 0.0280373832

<B さんが喜んでくれる A> 0.0280373832

<A を B 様に提供していく> 0.0280373832

<A を B 様のため> 0.0275229358

<B 様に A を提供する為> 0.0272727273

<B 様を飽きさせない A> 0.0272108844

<A は B 様に大好評です> 0.0260869565

<B 様に A をお届けできるよう> 0.0260869565

<B さまに最高の A> 0.0258620690

<B 様に A を提供していきます> 0.0250000000

<A を B 様にお届けしたい> 0.0248447205

<A で B 様を魅了する> 0.0240000000

<B 様は A を期待しています> 0.0235294118

<B 様に A をご提供> 0.0234375000

<A で B 様が喜ぶ> 0.0232558140

<B 様へ A をご提供すること> 0.0229885057

<A を重視する B 様> 0.0229007634

<A を B 様に提供できるように> 0.0229007634

<B さまに A をお届けすること> 0.0227272727

```
$ perl find-similar-patterns.pl ./patternDB.dat '<A は B が嫌い>' | head -30
```

<A は B が嫌いだ> 0.0318725100

<A は B が嫌いです> 0.0270522388

<A は B は嫌いです> 0.0239808153

<A は B が好きである> 0.0227479527

<A の好きなタイプの B> 0.0214285714

<A は B は嫌いだ> 0.0212290503

<A は B が嫌う> 0.0209481808

<A は B がとても好きです> 0.0208092486

<B が A は大好きです> 0.0207305035

<B が A は好きだ> 0.0196261682

<B が A は好きです> 0.0195048762

<A は B が大好きである> 0.0193003619

<A は B が大好きなのです> 0.0189165950

<B 嫌いの A にとる> 0.0184397163
<A は B が一番好きだ> 0.0178147268
<A は B が大嫌いです> 0.0177838577
<A は B がとても好きだ> 0.0176848875
<A は B が好きではありません> 0.0176565008
<B が好きな A とする> 0.0175781250
<A は B が嫌いなんだ> 0.0174672489
<B が苦手な A にとる> 0.0170031881
<B が好きではない A> 0.0166204986
<A は B が好きではない> 0.0165975104
<B を A は他に知らない> 0.0165413534
<A が B を好きになる> 0.0163934426
<A は B が好きなのだ> 0.0163511188
<A は B が好きになりました> 0.0163316583
<A も B が大好きです> 0.0160427807
<A も B が好きです> 0.0158862876

引数としてあげられる文字列が“<>”に囲まれていればスクリプトがそれをパターンと見なして、そのパターンの言い換えを検索します。スクリプトの引数としてファイルを指定することも可能です。ファイルですと、複数のパターンの言い換えを検索できます。ただし、その場合スクリプトが各パターンの言い換えを個別に計算するのではなく、ファイルに含まれる全パターンを一つのパターンと見なして、その全パターンの言い換えとなる表現を計算します。入力ファイルのフォーマットは１行・１パターンで、各パターンは “<…>”という形式が必要です。

4. 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。

5. 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