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はBの原因 | 0.0400322407 |
| Aの原因であるB | 0.0370898716 |
| AはBの原因にもなる | 0.0346598203 |
| AはBの原因となる | 0.0335473370 |
| Aの原因になる | 0.0328618063 |
| Bの原因にもなる | 0.0222953216 |
| AがBの原因となる | 0.0220117024 |
| AがBの原因である | 0.0206469374 |
| AがBの原因になる | 0.0204669359 |
| AはBを引き起こす | 0.0203133298 |
| Bの原因はA | 0.0197772221 |
| AがBの原因 | 0.0197204194 |
| AはBの原因 | 0.0191421482 |
| Aが原因で起こるB | 0.0191314910 |
| Aが原因となるB | 0.0186186186 |
| AがBを引き起こす | 0.0179789213 |
| AはBの原因である | 0.0175746924 |
| Bの原因 | 0.0173611111 |
| Bの原因となっている | 0.0169395822 |
| Bの原因は | 0.0166882463 |
| Aが引き起こすB | 0.0157584963 |

| A作とBの原因となる | 0.0155563071 |
| Aの原因とわかる | 0.0155361879 |
| AがBの原因である | 0.0137732223 |
| Bの原因のA | 0.0136869119 |
| BがA原因である | 0.0133979475 |
| Bの原因であるA | 0.0132726583 |
| Bの元となる | 0.0131449872 |
| AはBの原因になる | 0.0130890052 |
| Aを原因とするB | 0.0130201685 |
| Aによって起こるB | 0.0130054867 |
| AがBの原因であること | 0.0128544423 |
| Bの原因である | 0.0124058485 |
| AはBの原因 | 0.0123367199 |
| BがA原因 | 0.0122401848 |
| BのもととなるA | 0.0119910067 |
| AはBのもと | 0.0118072478 |
| AがBの原因となる | 0.0117379780 |
| Bの原因の一つであるA | 0.0117246596 |
| Bを引き起こすA | 0.0114091890 |
| Bの原因の一つであるA | 0.0108737864 |
| AがBの原因である | 0.0105860113 |
| ABの原因 | 0.0105288346 |
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。


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:

<p>&lt;パターン1&gt;  &lt;パターン2&gt; [コーパスでパターン1と共起する名詞対の数] [コーパスでパ
ターン2と共起する名詞対の数] [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
の数]

例:
&lt;A していることをB されました&gt;  &lt;A する可能性がB されています&gt; 40 103 2
&lt;A していることをB されました&gt;  &lt;A する可能性がB されています&gt; 40 177 4
&lt;A していることをB されました&gt;  &lt;A する可能性があることがB された&gt; 40 31 2
&lt;A していることをB されてきた&gt;  &lt;A する可能性もB されている&gt; 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。場所によって“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が多く含む 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 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
A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> 0.012363896
<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 予防する A> 0.0132532850
<B を防ぐ A です> 0.0118343195
<B で防止する 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
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
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