動詞含意関係データベース (Version 1.3.0)

目次

* 概要

このデータベースは、含意関係が成立している動詞のペアと含意関係が成立していないペアを列挙したものです。

動詞1が動詞2を含意するとは、動詞1の表す事態が成立するなら、同時かそれ以前に、動詞2の表す事態も成立しているということを意味します。例えば、「挑戦する」は「チャレンジする」を、「チンする」は「加熱する」を、「あざ笑う」は「笑う」を、「酔っ払う」は「飲む」を、「借りる」は「貸す」を含意します。

本データベースでは、含意が成立しているペア（正例）と成立していないペア（負例）をそれぞれ4種類に下位分類しています。正例負例の各下位分類とそのペア数、動詞1異なり数、動詞2異なり数を以下に示します。

1. 正例群
（ペア数:50,079　動詞1異なり数:24,948　動詞2異なり数:8,345）

1.1. 含意が成り立つ類義／上位下位関係
（ペア数:25,776　動詞1異なり数:13,814　動詞2異なり数:6,611）
動詞1と動詞2の間に含意が成立し、かつ、類義関係あるいは上位下位関係（動詞2が動詞1の上位概念）が成立している動詞ペアです。ただし、「含意が成り立つ類義／上位下位関係」の中には次に述べる「文字列上包含関係にあり、含意が成り立つ類義／上位下位関係」は含まれていません。
例）
挑戦する　チャレンジする
チンする　加熱する

1.2. 文字列上包含関係にあり、含意が成り立つ類義／上位下位関係
（ペア数:21,103　動詞1異なり数:16,835　動詞2異なり数:2,846）
含意が成り立つ類義／上位下位関係にあてはまる動詞ペアのうち、動詞1が動詞2を文字列上包含している動詞ペアを「文字列上包含関係にあり、含意が成り立つ類義／上位下位関係」と呼びます。
例）
あざ笑う　　笑う
セリーグ優勝する　リーグ優勝する

1.3. 前提関係
（ペア数:2,799　動詞1異なり数:2,190　動詞2異なり数:691）
動詞2が動詞1の前提条件になっている動詞ペアです。上の2種類の含意関係は動詞1の事態と動詞2の事態が同時に起こりますが、「前提関係」では、動詞2の

事態が動詞1の事態に時間的に先行します。
例）
酔っぱらう　飲む
稲刈する　　田植する

1.4. 作用反作用関係
（ペア数:401　動詞1異なり数:309　動詞2異なり数:308）
動作主体が異なる、一方が作用でもう一方が反作用と言える2つの動詞から成るペアです。一方、上の3種類の含意関係はいずれも、動詞1と動詞2の動作主体が同じです。
例）
借りる　貸す
受取る　手渡す

2. 負例群
（ペア数:38,787　動詞1異なり数:11,661　動詞2異なり数:4,612）

2.1. 含意、反義、予測関係ではない関連語ペア
（ペア数:38,239　動詞1異なり数:11,297　動詞2異なり数:4,509）
含意関係、あるいは以下で述べる反義関係、予測関係のいずれにも当てはまらないが、何らかの関連が認められるペアです。ただし、「含意、反義、予測関係ではない関連語ペア」の中には次に述べる「文字列上包含関係にあるが、含意、反義、予測関係ではない関連語ペア」は含まれていません。
例）
通勤する　走る
読書する　寛ぐ

2.2. 文字列上包含関係にあるが、含意、反義、予測関係ではない関連語ペア
（ペア数:373　動詞1異なり数:369　動詞2異なり数:157）
含意、反義、予測関係ではない関連語ペアのうち、動詞1が動詞2を文字列上包含している動詞ペアを「文字列上包含関係にあるが、含意、反義、予測関係ではない関連語ペア」と呼びます。
例）
冴渡る　　渡る
準優勝する　優勝する

2.3. 反義関係
（ペア数:24　動詞1異なり数:22　動詞2異なり数:16）
反義関係にあるペアです。
例）
閉める　　開ける
反比例する　比例する

2.4. 予測関係
（ペア数:151　動詞1異なり数:140　動詞2異なり数:111）
含意関係とは言えないが、動詞1の事態が起こるなら、その後動詞2の事態が起こる可能性が高いと言えるようなペアです。
例）
紅葉する　　落葉する
深煎りする　挽く

* 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、次の8つのファイルが含まれています。

ENT_ALAGIN_V1.3.0_entailment-ntriv.utf8
... 含意が成り立つ類義／上位下位関係

ENT_ALAGIN_V1.3.0_nonentailment-ntriv.utf8
... 含意、反義、予測関係ではない関連語ペア

ENT_ALAGIN_V1.3.0_entailment-triv.utf8
... 文字列上包含関係にあり、含意が成り立つ類義／上位下位関係

ENT_ALAGIN_V1.3.0_nonentailment-triv.utf8
... 文字列上包含関係にあるが、含意、反義、予測関係ではない関連語ペア

ENT_ALAGIN_V1.3.0_entailment-presu.utf8
... 前提関係

ENT_ALAGIN_V1.3.0_entailment-acrac.utf8
... 作用反作用関係

ENT_ALAGIN_V1.3.0_nonentailment-anton.utf8
... 反義関係

ENT_ALAGIN_V1.3.0_nonentailment-predi.utf8
... 予測関係

* ファイルフォーマット

上記8つのデータベースファイルはいずれも一行一ペアの形式になっています。
2つの動詞の間には半角スペースが介在しています。

正例データベース（含意が成り立つ類義／上位下位関係；文字列上包含関係にあり、含意が成り立つ類義／上位下位関係；前提関係；作用反作用関係）の場合、左側の動詞が右側の動詞を含意します。（作用反作用関係は動詞1から動詞2への含意関係だけでなく、動詞2から動詞1への含意関係も成立します。）次の例は含意が成り立つ類義／上位下位関係で、「履修する」が「学習する」、「在学する」、「受講する」、「勉強する」を含意することを表しています。

履修する 学習する
履修する 在学する
履修する 受講する
履修する 勉強する

負例データベース（含意、反義、予測関係ではない関連語ペア；文字列上包含関係にあるが、含意、反義、予測関係ではない関連語ペア；反義関係；予測関係）の場合、形式は正例データベースと同じですが、左側の動詞は右側の動詞を含意しません。含意、反義、予測関係ではない関連語ペアとして、例えば次のペアがあります。

通勤する 走る
通勤する 乗継ぐ
通勤する 乗換える
通勤する 通学する

各ファイルにおいて、サ変動詞（「する」で終わる動詞）を含むペアは、「する」無しのペアをその直後に追加しています。「する」無しのペアは「#」でマークされています。例えば次のようになっています。

履修する 学習する

履修 学習
履修する 在学する
履修 在学
履修する 受講する
履修 受講
履修する 勉強する
履修 勉強

また各ファイルにおいて、数字を含む動詞は、その数字の箇所を半角の「N」
あるいは「M」に置き換えています。例えば「NＲＫＯ勝ちする ＫＯ勝ちする」
「N曲プラスする N曲収録する」「N分のM削減する 減らす」などがあります。

* 作成方法

本データベースは、参考文献[1][2][4][5]の各手法で自動生成したものを人手で
アノテーションすることで作成しました。（参考文献[6]は[1]の内容を日本語化
したものです。）作成手順は次の通りです。

1. vt-名詞頻度テーブルの生成

大量のWeb文書を係り受け解析したデータ[3]から、

  名詞 助詞 動詞 （「野球 を 観戦する」など）

というタイプの係り受け関係を抽出し、「名詞 助詞 動詞」のインスタンス
ごとに頻度を計算して、次の形式のデータを得ます。

  名詞 助詞 動詞 頻度 （「野球 を 観戦する 418」など）

これをvt-名詞頻度テーブルと呼びます。助詞として、「は」「が」「を」「に」
「で」を使用しました。名詞と動詞には複合語も含まれます。また、否定形、
受け身形、使役形の動詞や出現頻度の低い動詞は対象から除外しています。
以下、「を 観戦する」等の「助詞 動詞」のまとまりをvtと表記します。

2. データベースの自動生成（正例、負例両者を含む）

上記1.で生成したvt-名詞頻度テーブルから、[1][2][4][5]の各手法に基づいて、
動詞ペアの含意スコアを計算します。この計算処理は、vt-名詞関連度テーブルの
生成と、それを用いた全vtペアの含意スコア計算（[1][2][4][5]の手法ごと）、
動詞ペアの含意スコア降順ソート（[1][2][4][5]の手法ごと）の3ステップで構成
されます。本データベースの正例と負例は、この自動生成結果を人手でアノテー
ションすることで得られました。

2.1. vt-名詞関連度テーブルの生成

vt-名詞頻度テーブルから vt-名詞関連度テーブルを生成します。[2][4][5]の
手法で用いるvt-名詞関連度テーブルは次の形式になっています。

```
 vt N:MI N:MI N:MI N:MI ...
 vt N:MI N:MI N:MI N:MI ...
 vt N:MI N:MI N:MI N:MI ...
   ...   ...   ...   ...
```

これは、元データにおいて共起している名詞（N）とvtとの間の相互情報量（MI）
を、vtごとに列挙したものです。ここでいう相互情報量は[2]に基づきます。vt
ごとに与えられた名詞群が、当該vtの文脈として利用されます。

一方、[1]の手法のvt-名詞関連度テーブルは次の形式になっています。

```
 vt N:P(vt|N):P(N|vt) N:P(vt|N):P(N|vt) N:P(vt|N):P(N|vt) ...
 vt N:P(vt|N):P(N|vt) N:P(vt|N):P(N|vt) N:P(vt|N):P(N|vt) ...
 vt N:P(vt|N):P(N|vt) N:P(vt|N):P(N|vt) N:P(vt|N):P(N|vt) ...
                ...    ...    ...    ...
```

これは、元データにおいて共起している名詞（N）とvtとの間の2種類の条件付き確率P(vt|N)とP(N|vt)を、vtごとに列挙したものです。vtごとに与えられた名詞群が、当該vtの文脈として利用されます。詳しくは[1]を参照してください。

2.2. 全vtペアの含意スコア計算

上記関連度テーブルを読み込み、[1][2][4][5]の手法別に、全 <vt1,vt2> ペア（ただし、vt1≠vt2）の含意スコアを計算します。以下で、各手法の計算結果を[1].out、[2].out、[4].out、[5].outと呼びます。手法[4]に関しては、さらに、複合名詞／動詞の認定基準と出現頻度閾値が異なるもう一つのvt-名詞頻度テーブルを用いて <vt1,vt2> ペアの含意スコアを計算しました。この結果を以後[4].out2と呼びます。

2.3. 動詞ペアの含意スコア降順ソート

計算結果ごとに含意スコアの降順にペアをソートします。その後、各vtから助詞の部分を取り除き、vtペアの含意リストから動詞ペアの含意リストへと変換します。

3. データベースの人手アノテーション

ソートされた[1].out、[2].out、[4].out、[5].out、[4].out2を人手で正例／負例のアノテーションをします。ただし、これら全てをアノテーションするのではなく、次の2つの条件A)とB)のいずれかに合致するものだけを対象にします。

A) [1].out、[2].out、[4].out、[5].outのスコア上位50,000位内と、[4].out2のスコア上位37,000位以内にある動詞ペア。

B) [1].out、[2].out、[4].out、[5].outのスコア上位1,000,000位内にある動詞ペア「v1 v2」のうち、v1、v2の両方が形態素解析器JUMANの辞書に登録されていて、かつ、動詞含意関係データベースのVersion 1.1.1に次の条件を満たすペアが存在するもの：含意される側の動詞がv2と同じで、含意する側の動詞がv1をsuffixとして持つ。例えば、JUMAN辞書に登録されている動詞から成るペア「ソテーする いためる」は[1].outの上位1,000,000位内に存在し、かつ、動詞含意関係データベース version 1.1.1には「バターソテーする いためる」というペアがあるので、本条件B)に合致します。

アノテーション前に、あらかじめ、次のいずれかに該当するペアは除外しました。

- 猥褻語を含むペア
- 末尾が「〜て参る」「〜ならない」「〜て下さる」等の動詞を含むペア

アノテーションでは、各ペアに対して次のいずれかのマークを付与しました。

- o: 正例
- x: 負例
- ?: 動詞の意味が不明で判定不能

4. アノテーション結果のマージと正例、負例の下位分類

全アノテーション結果を正例、負例ごとにマージした上で、正例を、含意が

成り立つ類義／上位下位関係、文字列上包含関係にあり、含意が成り立つ
類義／上位下位関係、前提関係、作用反作用関係のいずれかに分類します。
同じく、アノテーション結果の負例を、含意、反義、予測関係ではない関連語ペア、
文字列上包含関係にあるが、含意、反義、予測関係ではない関連語ペア、反義関係、
予測関係のいずれかに分類します。最後に、下位分類ごとに、動詞2の辞書順に
ソートします。

* 変更履歴

** Version 1.2.0からの変更点

- 正例、負例ともに下位分類を設けました。それに伴い、正例を29,458ペアから
50,079ペアに、負例を38,610ペアから38,787ペアに増強しました。
- Version 1.2.0の正例／負例の判定誤りを修正しました。

** Version 1.1.1からの変更点

- 正例約5,500ペア（[2]と[4]のスコア上位5万以内の動詞ペアをアノテーション
したものと、[1][2][4][5]のスコア上位100万以内にあるJUMAN動詞ペア
（JUMANの辞書に登録されている動詞から成るペア）をアノテーションしたもの）
を追加しました。
- 正例／負例の判定の誤り（約20ペア）を修正しました。

** Version 1.1からの変更点

- [1]と[5]による自動獲得結果をアノテーションしたもの（負例）を追加しました。

** Version 1.0からの変更点

- [1]と[5]による自動獲得結果をアノテーションしたもの（正例のみ）を追加
しました。
- Version 1.0の正例(ENT_ALAGIN_V1_entailment.utf8)のファイル形式
は、動詞1でソートしたものですが、Version 1.1の正例
(ENT_ALAGIN_V1_entailment.utf8)は、動詞2でソートしています。

* 利用条件

本データベースの利用には、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] Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Masaki Murata, and Jun'ichi Kazama. Large-Scale Verb Entailment Acquisition from the Web. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2009), pages 1172-1181. 2009.

[2] Dekang Lin and Patrick Pantel. Discovery of Inference Rules for Question Answering. Natural Language Engineering, Vol.7, Num.4, pages 343-360. 2001.

[3] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. Tsubaki: An open search engine infrastructure for developing new information access.
In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP2008), pages 189-196, 2008.

[4] Idan Szpektor and Ido Dagan. Learning Entailment Rules for Unary Templates. In Proceedings of the 22nd International Conference on Computational Linguistics (COLING2008), pages 849-856, 2008.

[5] Julie Weeds and David Weir. A general framework for distributional similarity. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2003), pages 81-88. 2003.

[6] 橋本力, 鳥澤健太郎, 黒田航, デサーガ・ステイン, 村田真樹, 風間淳一.
WWWからの大規模動詞含意知識の獲得. 情報処理学会論文誌,
Volume 52 Number 1, pp.293--307. 2011.

* 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