基本的意味関係の事例ベース (Version 1.0)

Created: 2010/05/28
Updated: 2010/06/07, 09, 11

目 次

はじめに

0.1 公開履歴

(1) 2010 年 06 月 16 日: Version 1.0 公開

a. 同義語句対 [s] の数: n

b. 略語対 [a] の数: n

c. 対義語句対 [s] の数: n

d. 部分・全体語句対 [p] の数: 1,318

0.2 ファイルの一覧

- classified-pairs-v1.sjis.csv.zip [文字コード Shift-JIS (Windows 環境向け)]
- classified-pairs-v1.eucj.csv.zip [文字コード EUC-JP (Unix/Linux 環境向け)]
- classified-pairs-v1.utf8.csv.zip [文字コード UTF-8 (汎用)]

0.3 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.

0.4　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．

0.5　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

以下，1 節ではデータの基本的な特徴を説明し．2 節でデータ構造と付与情報の詳細を説明する．3 節でデータの構築法の詳細を解説します．

1　データの簡単な解説

1.1　データの見本

同義語句対 (分類ラベル s)，略語対 (分類ラベル a)，対義語句対 (分類ラベル d)，部分・全体語句対 (分類ラベル p) の例をそれぞれ，(1)，(2)，(4)，(3) に示します (表示形式は分類ラベルで始まる，コンマ区切り形式 (comma-separated value: csv) で，ファイルにある通りです):

(1)　i. s, ＡＤＳＬ接続サービス, ＡＤＳＬサービス

ii. s, 除細動器, ＡＥＤ

iii. s, 狂牛病, ＢＳＥ

(2)　i. a, 国連, 国際連合

ii. a, メーリングリスト,ML

iii. a,SE, システムエンジニア

(3)　i. d, 夜, 朝

ii. d, 免税事業者, 課税事業者

iii. d, 後半, 前半

(4)　i. p, 学部, 大学

ii. p, 東南アジア, アジア

iii. p, 清朝末期, 清代

注意: 部分・全体の関係は非対称であり，語句の出現順序に依存する．データでは “p, 部分を表わす語句, 全体を表わす語句” としています.

行の先頭にある一文字のアルファベット (例えば s, a, p, d) が分類のタイプです．これらのタイプの詳細は §2.1 で説明します.

1.2　本データの典型的な使い方

本データは汎用的なもので，用途は特定のものに限定はされませんが，典型的には次のような使い方が考えられます．

同義語対と略語対は Web 検索での検索式の拡張に有用です．(狂牛病, BSE), (狂牛病, MCD) のような対があれば，「狂牛病」の検索式を「狂牛病 or BSE or MCD」に拡張できます．ただし，追加する語句によっては曖昧性が増しすぎる場合があるので，注意が必要です．

部分・全体語対は省略解析を含む推論に有用なデータです．部分から全体を推測する場合であれば，例えば，(拡張子, ファイル名) という対から，拡張子がファイル名の一部であることがわかり，その上で ⟨ ファイル名 ⟩ が ⟨ ファイル ⟩ の属性だとわかれば，それを使って「その時には拡張子を変えて開いた」という表現から，開かれたのがファイルだったことが推測できます (*X* が個体であれば，⟨*X* 名 ⟩ が ⟨*X*⟩ の属性であるのはかなり一般的な規則なので，二番目の段階の推論はそれほど難しくありません)．全体から部分を推測することも可能ですが，部分と全体の関係には非対称性があり，こちらは部分から全体を推測するよりは精度が下がる可能性があります．

2　データ構築の概要

2.1　作成方法

本データは黒田ら (2010) が報告した異表記認識のための分類基準の拡張版 (未公開) に基づいて，次の手順で作成されました:

(5) 手順

Step 1: 風間ら (2009) の手法で構築された名詞句のクラスター化データを基にして，見出し語句 w0 とそれに文脈類似度が最大の語句 w1 と第 2 位の語句 w2 を選び，(w0, w1) と (w0, w2) という対を生成する．w0 としては，成語性のない文字列や定型的な

パターンをもった一部の語句を除いた上位 15 万とした．これにより，30 万対の評定の候補が生成される．

Step 2: こうして生成された 30 万個の対のそれぞれを，§2.2 の (6) に示す 18 個の基準 [s, n, a, v, e, f, m, h, p, k, w, c, d, t, o, u, x, y] で人手分類した．

Step 3: その分類の結果を，ラベルごとに人手で最終チェックした．

ただし，公開されるのはこのように評定されたデータの一部のみ．

2.2 分類基準

(6) 人手分類の基準

s: 同義異語句対: 同じ対象を指示する (ことのある) 異なる語句の対である場合．例えば
[用紙トレイ, 給紙トレイ], [学園闘争, 学園紛争], [単独首位, 単独トップ], [パイプ内, 配管内], [ガウス分布, 正規分布], [買い手, 売る相手], [責任逃れ, 言い逃れ]

a: 略語対: 同じ語句の異なる表記の対だが，一方が他方の略式表記になっている場合．例えば
[慶応大学, 慶大], [短期大学, 短大], [HDD, ハードディスクドライブ]

n: 条件つき異名対: 一方が多方の「あだ名」や「値」になっている場合．例えば
[人間機関車, 浅沼稲次郎], [アメリカ大統領, バラク・オバマ], [安倍首相, 安倍元首相]

v: 同語異表記対: [a] を除いて同じ語句の異なる表記の対である場合．例えば
[一リーグ制, 1 リーグ制], [100 メートル, 100m], [57 キロ, 57k], [ハンナ・アーレント, ハンナ・アレント], [オーソリティ, オーソ

リティー], [憂鬱, ゆううつ], [肩掛け，肩かけ], [アタリ, ATARI], [Kernel, kernel], [PHPMySQL, PHP MySQL], [お問い合わせ, 問合せ], [海ヘビ, うみへび]

e: 誤表記対: v の特殊な場合で，一方が他方の誤表記だと判断できる場合．例えば
[メールアドレス, ルアドレス], [もらい手, らい手], [シミュレーション, シミュュレーション]

f: 準誤表記対: 本来は誤記だと思われる表記が正用化していると判断できる場合．例えば
[サンドバッグ, サンドバック] (cf. バック転 vs *バッグ転), [シミュレーション, シュミレーション]

m: 誤用対: s の特殊な場合で，異なる語句が変換ミスなどによって偶発的に同じ意味で使われていると判断できる場合．例えば
[精算金, 清算金], [化学兵器, 科学兵器]

h: 上位語と下位語の対．例えば
[柴犬, 犬], [再婚, 結婚]

p: 部分を表わす語句と全体を表わす語句との対．例えば
[太平洋戦争, 第二次世界大戦], [椅子, 背もたれ], [ジョン・レノン, ビートルズ]

k: 過度に抽象的でない共通の上位語をもつ同類語で，形態素共有のない語句の対．例えば
[タイ, アルゼンチン], [イワシ, サンマ],

w: 同類語句対のうち，形態素共有のある場合 (共通の上位語をもつ同類語で (主に語句末で) 形態素を共有する)．例えば
[中国, 韓国], [二日, 三日], [土曜日, 日曜日]

c: 対比性をもつ語句の対 (これは同類語 [k] の特殊な場合)．例えば

[ジョン・レノン，ポール・マッカートニー] [リール, 釣竿], [エンジン, タイヤ]

d: 対義性をもつ語句の対 (これは対比語句対 [c] の特殊な場合). 例えば
[右側, 左側], [高抵抗, 低抵抗]

t: 時間上の順序づけが可能な語句の対．例えば
[離婚, 結婚], [再婚, 離婚], [出産, 妊娠]

o: 語句対に関連性はあるが，それが上記の s, a, n, v, e, f, m, h, p, k, w, c, d, t のいずれでもない場合

u: 無関連語対: 両方の語句が意味をなすが，はっきりと認識できる関連性があると認識できない場合．例えば
[風習, アーム], [船体, 仙臺]

x: 無意味語対: 少なくとも一方が意味をなさない語句である場合. 例えば
[い出, 思い出], [もら, もち]

y: 評定不能

参照文献

- 風間 淳一, De Saeger, S., 鳥澤 健太郎 and 村田 真樹 (2009). 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. In **言語処理学会第第 15 回年次大会発表論文集**, pp. 84–87.

- 黒田 航, 風間 淳一, 村田 真樹 and 鳥澤 健太郎 (2010). Web データに対応できる日本語異表記対の認定基準. In **言語処理学会 16 回年次大会発表論文集**, pp. 990–993.