基本的意味関係の事例ベース付録: 用語抽出用評価データ

1. 公開履歴

2010 年 09 月 26 日　　公開

2. 「用語抽出用評価データ」について

2.1　　概要

本データは、用語抽出の評価データとして役立つように構築されたものです。このデータでは、用語データを極力漏れの少ない形で作成しております。用語抽出の実験において、再現率・
適合率を算出するなどの性能評価に用いることが可能です。なお、データは SJIS で記載しております。

2.2　　データの種類

一項のデータと二項データがあります。一項データの一例には国名があり、国名のリストの場合は
以下のようなデータが収録されています。

アイスランド アイスランド共和国 ISL
アイルランド アイルランド共和国 IRL
アゼルバイジャン アゼルバイジャン共和国 AZE
アゾレス諸島
...

一項データには国名以外に、「太陽系惑星」など合計 58 種類のリストがあります。異表記も記載しております。

二項データの一例には国名と首都の対があり、国名と首都の対の場合は以下のデータがあります。

アイスランド アイスランド共和国 ISL レイキャビク レイキャヴィーク レイキャヴィク
アイルランド アイルランド共和国 IRL ダブリン
アゼルバイジャン アゼルバイジャン共和国 AZE バクー
アゾレス諸島 アングラ
...

二項データには国名と首都名の対以外に、「太陽系惑星−衛星」など合計 58 種類があります。

2.3 ディレクトリの説明

D_SINGLE --- 一項データのファイルには、58 種類のデータ分のファイルがあります。種類ごとに別ファイルにデータを格納しております。

D_PAIR --- 二項データのファイルには、58 種類のデータ分のファイルがあります。種類ごとに別ファイルにデータを格納しております。

LIST --- データ作成者による補足説明を記載しています。

LIST_SINGLE.txt は一項データの補足説明を記載しています。

LIST_PAIR.txt は二項データの補足説明を記載しています。

これらのファイルを見ることで 58 種類がどのようなものであるかを把握することができます。

2.4 LIST ファイルの見かた

LIST ディレクトリにある LIST_SINGLE.txt や LIST_PAIR.txt は1行が1種類のデータに対応し、タブ区切りで表現しています。データは、タブ区切りで、左から順に、データ番号、単独データ（データの種類の名称）、ファイル、代表形基準コメント、異表記基準コメント、その他コメント、数、網羅度を表記しています。

ファイルは、D_SINGLE または D_PAIR の下にある、そのデータのファイル名です。数はその種類のデータが何個あるかを意味します。網羅度は、その種類のデータをどのくらい網羅して収集できたかを意味します。○はほぼ 100%網羅している，△と×は網羅していないことを意味します。例えば国名のようにある時点に限ればほぼ 100%網羅したデータを作成可能であり、○はそのようなデータを意味します。100%網羅したデータのため、用語抽出の際の再現率・適合率を算出することが可能となります。

2.5 データの見かた

D_SINGLE または D_PAIR の下にデータを記載したファイルがあります。そのファイル内でのデータの表示方法を以下に記載します。1行が1データであり、各行に異表記をタブ区切りで記載しています。

D_PAIR の下にある二項のデータでは、二項のものを//で区切って表記しております。例えば、

アイスランド	アイスランド共和国	ISL	//	レイキャビク	レイキャヴィーク
アイルランド	アイルランド共和国	IRL	//	ダブリン	

のように、国名と首都名を//で区切って表記しています。

3. 「基本的意味関係の事例ベース付録:用語抽出用評価データ」ご利用に際して

利用許諾契約

「基本的意味関係の事例ベース付録:用語抽出用評価データ」(以下、「本データベース」と呼ぶ)の利用には、独立行政法人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。

なお、すでに「基本的意味関係の事例ベース Version 1」をご利用いただいている方は、新たに利用許諾契約を行うことなく「用語抽出用評価データ」をご利用いただけます。

利用上の注意

本データベースは、インターネットホームページ等、独立行政法人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、独立行政法人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、独立行政法人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。

問い合わせ先

独立行政法人情報通信研究機構 ユニバーサルコミュニケーション研究所

情報分析研究室

Email: alagin-lr@khn.nict.go.jp

参考文献

村田 真樹, 馬 青, 白土 保, 井佐原 均. 用語抽出用評価データの作成とその利用

言語処理学会 第10回年次大会併設ワークショップ「固有表現と専門用語」 2004 年. 東京工業大学.