Japanese WordNet Synonyms Database

ver.1.0

1. Database overview

Japanese WordNet Synonyms Database is a collection of 11,753 synonym pairs, which were collected using synsets in Japanese WordNet version 1.1. Word pairs were created using words in a synset, which is a cluster of words that share the same sense, and were manually annotated. The word pairs that were manually annotated as synonym pairs were included in the database. For instance, under synset 00623862-n, there are following words.

izakoza, komarimono, wazawai, gokuro, kosho, konnan, haran, moyakuya, isakusa, toraburu

From this synset, word pairs such as in the ones in the followings were created.

izakoza, komarimono

izakoza, wazawai

izakoza, gokuroo

izakoza, kosho

izakoza, konnan

…

From the word pairs, only the ones that were manually annotated as a synonym pair were included in the database. Synonym pairs that are in a linguistic resource of ALAGIN forum, and the ones of which frequency is less than 1,500 times in a corpus composed of about 600,000,000 webpages were excluded from the manual annotation.

2. Recognition criterion for synonym pairs

In this database, if Word A in a given clause can be replaced with Word B, maintaining the same meaning, a pair of Word A and Word B are annotated as a synonym pairs. For instance, the term *gohan* can be replaced with the term *meshi*, in the clause “*gohan-o taberu*” (I am about to eat cocked rice). Based on the fact that *gohan* and *meshi* can be replaced, they are annotated as a synonym pair. In addition, if word pairs have the relationships described in Table 1, they are also included in the database.

Relationship between words	Example
notational variation	お米，おこめ/サーバー，サーバ
abbreviation	メールアドレス,メアド
formal – informal	御本,本/高覧,見る/お母さん,母
dialectal relation	かわず,カエル
everyday – specialised	リンドウ，ゲンティアナ・ベルナ
Japanese words – loan words	案内，ガイダンス
taditional naming – modern naming	江戸,東京
alias	ラフカディオハーン,小泉八雲
metaphor	犬,スパイ

Table 1： list of relationship between words that were included in the database

3. Format

The fields in the database are described in Table 2.

Number of the fields	Name of the fields	Description
1	word1ID	Word ID used in Japanese WordNet. If there are more than one ID for one word, it is marked with "@"
2	word1	Lemma which has word1ID
3	word2ID	Word ID used in Japanese WordNet. If there are more than one ID for one word, it is marked with "@"
4	word2	Lemma which has word2ID

Table 2： Explanation of the fields in the database

Examples are provided below.

217548	いさかい	190577	口げんか
157025	うす茶色	228576	ライトブラウン
248196	おしまい	199808	終了
232728	ほうき星	183633	コメット
178252	やや	225897	赤ちゃん
207752	やり方	187525	スタイル
182109	イラスト	222835	挿絵
179506	クラス	202454	科目
179506	クラス	219478	等級
174913	シネマ	248877	映画
199908	上天気	207086	晴天

4. License

The database is available under the license described below.

Copyright: 2009, 2010

This software and database is being provided to you, the LICENSEE, by the National Institute of Information and Communications Technology under the following license. By obtaining, using and/or copying this software and database, you agree that you have read, understood, and will comply with these terms and conditions:

> Permission to use, copy, modify and distribute this software and database and its documentation for any purpose and without fee or royalty is hereby granted, provided that you agree to comply with the following copyright notice and statements, including the disclaimer, and that the same appear on ALL copies of the software, database and documentation, including modifications that you make for internal use or for distribution.

Japanese WordNet Copyright 2009, 2010 by the National Institute of Information and Communications Technology (NICT). All rights reserved.

THIS SOFTWARE AND DATABASE IS PROVIDED "AS IS" AND NICT MAKES NO REPRESENTATIONS OR WARRANTIES, EXPRESS OR IMPLIED. BY WAY OF EXAMPLE, BUT NOT LIMITATION, NICT MAKES NO REPRESENTATIONS OR WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE OR THAT THE USE OF THE LICENSED SOFTWARE, DATABASE OR DOCUMENTATION WILL NOT INFRINGE ANY THIRD PARTY PATENTS, COPYRIGHTS, TRADEMARKS OR OTHER RIGHTS.

The name of the National Institute of Information and Communications Technology may not be used in advertising or publicity pertaining to distribution of the software and/or database. Title to copyright in this software, database and any associated documentation shall at all times remain with National Institute of Information and Communications Technology and LICENSEE agrees to preserve same.

5. Contact

Information Analysis Laboratory, Universal Communication Research Institute, National Institute of Information and Communications Technology.

Please contact the e-mail address below for enquiry.

Email: jwordnet@gmail.com