Linking and Extending an Open Multilingual Wordnet

Francis Bond
Linguistics and Multilingual Studies
Nanyang Technological University
bond@ieee.org

Ryan Foster
Great Achievement Press
ryan_foster@gapbks.com

Abstract

We create an open multilingual wordnet with large wordnets for over 26 languages and smaller ones for 57 languages. It is made by combining wordnets with open licences, data from Wiktionary and the Unicode Common Locale Data Repository. Overall there are over 2 million senses for over 100 thousand concepts, linking over 1.4 million words in hundreds of languages.

1 Introduction

We wish to create a lexicon covering as many languages as possible, with as much useful information as possible. Generally, language resources, to be useful, must be both **accessible** (legal to use) and **usable** (of sufficient quality, size and with a documented interface) (Ishida, 2006). We address both of these concerns in this paper.

One of the many attractions of the semantic network WordNet (Fellbaum, 1998), is that there are numerous wordnets being built for different languages. There are, in addition, many projects for groups of languages: Euro WordNet (Vossen, 1998), BalkaNet (Tufiş et al., 2004), Asian Wordnet (Charoenporn et al., 2008) and more. Although there are over 60 languages for which wordnets exist in some state of development (Fellbaum and Vossen, 2012, 316), less than half of these have released any data, and for those that have, the data is often not freely accessible (Bond and Paik, 2012). For those wordnets that are available, they are of widely varying size and quality, both in terms of accuracy and richness. Further, there is very little standardization in terms of format, what information is included, or license.

The goal of the research outlined in this paper is to make it possible for a researcher interested in working on the lexical semantics of a language or languages to be able to access wordnets for those languages with a minimum of legal and technical barriers. In practice this means making it possible to access multiple wordnets with a common interface. We also use sources of semi-structured data that have minimal legal restrictions to automatically extend existing freely available wordnets and to create additional wordnets which can be added to our open wordnet grid.

Previous studies have leveraged multiple wordnets and Wiktionary (Wikimedia, 2013) to extend existing wordnets or create new ones (de Melo and Weikum, 2009; Hanoka and Sagot, 2012). These studies passed over the valuable sense groupings of translations within Wiktionary and merely used Wiktionary as a source of translations that were not disambiguated according to sense. The present study built and extended wordnets by directly linking Wiktionary senses to WordNet senses.

Meyer and Gurevych (2011) demonstrated the ability to automatically identify many matching senses in Wiktionary and WordNet based on the similarity of monolingual features. Our study combines monolingual features with the disambiguating power of multiple languages. In addition to differences in linking methodology, our project gives special attention to ensuring the maximum re-usability and accessibility of the data and software released.

Other large scale multilingual lexicons have been made by linking wordnet to Wikipedia (Wikipedia, 2013; de Melo and Weikum, 2010; Navigli and Ponzetto, 2012). Our approach is complementary to these: in general Wikipedia has more entities than classes, while Wiktionary has more classes.

In Section 2 we discuss linking freely available wordnets to form a single multilingual semantic network. In Section 3 we extend the wordnets with data from two sources. We show the results in Section 4 and then discuss them and outline future

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1352–1362,
Sofia, Bulgaria, August 4-9 2013.

work in Section 5.

2 Linking Multiple Wordnets

In order to make the data from existing wordnet projects more **accessible**, we have built a simple database with information from those wordnets with licenses that allow redistribution of the data. These wordnets, their licenses and recent activity are summarized in Table 1 (sizes for most of them are shown in Table 2).[1]

Wordnet Project	Lng	Licence	Type
Albanet^{o}	als	CC BY	a
Arabic WordNet	arb	CC BY-SA	s
DanNet	dan	wordnet	a
Princeton WordNet^{u}	eng	wordnet	a
Persian Wordnet	fas	free to use	u
FinnWordNet^{u}	fin	CC BY	a
WOLF^{u}	fra	CeCILL-C	s
Hebrew Wordnet^{o}	heb	wordnet	s
MultiWordNet^{o}	ita	CC BY	a
Japanese Wordnet^{u}	jpn	wordnet	a
Multilingual	cat	CC BY	a
Central	eus	CC BY-NC-SA	n
Repository^{o,u}	glg	CC BY	a
	spa	CC BY	a
Wordnet Bahasa^{u}	ind	MIT	a
	zsm	MIT	a
Norwegian Wordnet^{o}	nno	wordnet	a
	nob	wordnet	a
plWordNet^{o,u}	pol	wordnet	a
OpenWN-PT^{u}	por	CC BY-SA	s
Thai Wordnet	tha	wordnet	a

^{o} Re-released under an open license in 2012
^{u} Updated in 2012
Type: **u** Unrestricted; **a** Attribution; **s** Share-alike; **n** Non-commercial
URL: `http://casta-net.jp/~kuribayashi/multi/`

Table 1: Linked Open Wordnets

The first wordnet developed is the Princeton WordNet (PWN: Fellbaum, 1998). It is a large lexical database of English. Open class words (nouns, verbs, adjectives and adverbs) are grouped into concepts represented by sets of synonyms (synsets). Synsets are linked by semantic relations such as hyponomy and meronomy. PWN is released under an open license (allowing one to use, copy, modify and distribute it so long as you properly acknowledge the copyright).

The majority of freely available wordnets take the basic structure of the PWN and add new lemmas (words) to the existing synsets: the **extend** model (Vossen, 2005). For example, $\mathsf{dog}_{\text{n:1}}$ is linked to the lemmas *chien* in French, *anjing* in Malay, and so on. It is widely realized that this model is imperfect as different languages lexicalize different concepts and link them in different ways (Fellbaum and Vossen, 2012). Nevertheless, many projects have found that the overall structure of PWN serves as a useful scaffold. The fact that, for example, a $\mathsf{dog}_{\text{n:1}}$ is an $\mathsf{animal}_{\text{n:1}}$ is language independent.

In theory, such wordnets can easily be combined into a single resource by using the PWN synsets as pivots. All languages are linked through the English wordnet. Because they are linked at the synset level, the problem of ambiguity one gets when linking bilingual dictionaries through a common language is resolved: we are linking senses to senses.

In practice, linking a new language's wordnet into the grid could be problematic for three reasons. The first problem was that the wordnets were linked to various versions of the Princeton WordNet. In order to combine them into a single multilingual structure, we had to map to a common version. The second problem was the incredible variety of formats that the wordnets are distributed in. Almost every project uses a different format. Even different versions of the same project often had slightly different formats. The final problem was legal: not all wordnets have been released under licenses that allow reuse.

The first problem can largely be overcome using the mapping scripts from Daude et al. (2003). Mapping introduces some distortions, in particular, when a synset is split, we chose to only map the translations to the most probable mapping, so some new synsets will have no translations.

The second problem we are currently solving through brute force, writing a new script for every new project we add. We make these scripts, along with the reformatted wordnets, freely available for download. Any problems or bugs found when converting the wordnets have been reported back to the original projects, with many of them fixed in newer releases. We consider this feedback to be an important part of our work: it means that other researchers and users do not have to suffer from the same problems and it encourages projects to release updates.

The third, legal, problem is being solved by an ongoing campaign to encourage projects to (re-)release their data under open licenses. Since Bond and Paik (2012) surveyed wordnet licenses in 2011, six projects have newly released data un-

[1] We have now added Mandarin Chinese.

der open licenses and eight projects have updated their data.

Our combined wordnet includes English (Fellbaum, 1998); Albanian (Ruci, 2008); Arabic (Black et al., 2006); Chinese (Huang et al., 2010); Danish (Pedersen et al., 2009); Finnish (Lindén and Carlson., 2010); French (Sagot and Fišer, 2008); Hebrew (Ordan and Wintner, 2007); Indonesian and Malaysian (Nurril Hirfana et al., 2011); Italian (Pianta et al., 2002); Japanese (Isahara et al., 2008); Norwegian (Bokmål and Nynorsk: Lars Nygaard 2012, p.c.); Persian (Montazery and Faili, 2010); Portuguese (de Paiva and Rademaker, 2012); Polish (Piasecki et al., 2009); Thai (Thoongsup et al., 2009) and Basque, Catalan, Galician and Spanish from the Multilingual Common Repository (Gonzalez-Agirre et al., 2012).

On our server, the wordnets are all in a shared `sqlite` database using the schema produced by the Japanese WordNet project (Isahara et al., 2008). The database is based on the logical structure of the Princeton WordNet, with an additional language attribute for lemmas, examples, definitions and senses. It is a single open multilingual resource. When we redistribute the data, each project's data is made available separately, with a common format, but separate licenses.

The Scandinavian and Polish wordnets are based on the **merge** approach, where independent language specific structures are built and then some synsets linked to PWN. Typically only a small subset will be linked (due more to resource limitations than semantic incompatibility).

2.1 Core Concepts

Boyd-Graber et al. (2006) created a list of 5,000 **core** word senses in Princeton WordNet which represent approximately the 5,000 most frequently used word senses.[2] We use this list to evaluate the coverage of the wordnets: do they contain words for the most common concepts? As a very rough measure of useful coverage, we report the percentage of synsets covered from this core list. Because the list is based on English data, it is of course not a perfect measure for other languages and cultures. Note that some wordnet projects have deliberately targeted the core concepts, which of course boosts their coverage scores.

[2]The original list is here from `http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt`; we converted it to wn30 synsets.

2.2 License Types

The licenses fall into four broad categories: (**u**) completely unrestricted, (**a**) attribution required, (**s**) share alike, and (**n**) non-commercial. The first category includes any work that is in the public domain or that the author has released without any restrictions. The second category allows anyone to use, adapt, improve, and redistribute the work as long as one attributes the work in the manner specified by the copyright holder (without suggesting an endorsement). The WordNet, MIT, and CC BY licenses are all in this category. The third category allows anyone to adapt and improve the licensed work and redistribute it, but the redistributed work must be released under the same license. The CC BY-SA, GPL, GFDL, and CeCILL-C licenses are of this type. Because derivative works can only be redistributed under the same license, works licensed under any two of these licenses cannot be combined with each other and legally redistributed. In general, a work formed from the combination of works in category (**u**) and (**a**) with a work in category (**s**) will be subject to the more restrictive terms of the the share alike license. However, the GPL, GFDL and CeCILL-C are incompatible with CC BY.[3] The fourth type of license further forbids the commercial use of a work. The CC BY-NC and the CC BY-NC-SA licenses are in this category, they are also incompatible with licenses in category (**s**).

Releasing a work under the more restrictive licenses in categories (**s**) and (**n**) above substantially limit and complicate the ability to extend and combine a work into other useful forms. By maintaining a separation of databases released under incompatible licenses, we avoid any possible legal problems. Due to license incompatibilities, it is impossible to release a single database with all the wordnets, even though individually they are redistributable. We can currently combine those with licenses in groups (**u**) and (**a**) and the CC BY-SA wordnets (now everything except French and Basque).

3 Extending with non-wordnet data

We looked at two sources for automatically adding new entries. The Unicode Common Locale Data Repository (CLDR) has reliable information on languages, territories and dates. Wiktionary is a

[3]`http://www.gnu.org/licenses/license-list.html\#ccby`

general purpose lexicon with much more information for many words.

3.1 Unicode Common Locale Data Repository (CLDR)

We added information on languages, territories and dates from the Unicode Common Locale Data Repository (CLDR).[4] This is a collection of data maintained by the Unicode Consortium to support software internationalization and localization with locale information on formatting dates, numbers, currencies, times, and time zones, as well as help for choosing languages and countries by name. It has this data for over 194 languages. It is released under an open license that allows redistribution with proper attribution (Unicode, Inc., 2012).[5]

We found data for 122 languages. Most had around 550 senses (synsets and their lemmas): for example, for Portuguese: English$_{n:1}$ *inglês*. Some had only 40 or 50, such as Assamese, which only has the week days, month names and a few language names. The linked data was small enough to check by hand. When the original CLDR data is correct the data we generate should be correct.

The idea of using such data is not new. Quah et al. (2001) for example, use Linux locale data to extend a proprietary English-Malay lexicon. de Melo and Weikum (2009) also use this data (and data from a variety of other sources) to build an enhanced wordnet, in addition adding new synsets for concepts that are not in not wordnet. However, when they released the data as LEXVO (data about languages: CC BY-SA) and UWN (the universal multilingual wordnet: CC BY-NC-SA), they added additional license restrictions which complicate the reuse of the data and make it impossible to integrate the data back into the original wordnet project.

3.2 Wiktionary

Searches for a publicly-available source of Wiktionary in a preprocessed, machine-readable format did not turn up any sources that were recent and publicly-available.[6] Although there are several freely-available software programs that are capable of parsing portions of the English Wiktionary, none of the programs that were evaluated appeared to extract the precise set of information desired for our task in an easy-to-use format. So the authors decided to build a custom parser capable of extracting the information needed for building open wordnets.

3.2.1 Wiktionary Parser

Since each language edition of Wiktionary is formatted in a somewhat unique way, parsers must be tailored to recognize the structure and formatting of each edition on a case-by-case basis. The authors created a parser tailored to the English Wiktionary, although it can be extended to handle other language versions as well. We are releasing this code under the MIT license.[7]

The current version of the parser is capable of extracting headwords, parts of speech, definitions, synonyms and translations from the XML Wiktionary database dumps provided by the Wikimedia Foundation.[8] Within these large XML files, the main body of Wiktionary articles are stored in a Wikitext format, which is a semi-structured format. Although anyone can edit a Wiktionary page and use any style of formatting they desire, the community of users encourages adhering to established guidelines, which produces a format that is generally predictable.

Within the English Wiktionary, synonyms and translations are both grouped into sense groups that correspond with definitions in the main section. These sense groups are marked by a short text gloss (**short gloss**), which is usually an abbreviated version of one of the full definitions (**full definition**). The parser makes no attempt to match these short glosses with the full definitions. Data is simply extracted, cleaned, and then stored in a relational database or flat file.

Translations proved to be easy to extract due to the fairly consistent use of a specifically formatted translation template. These templates include a language code derived from ISO standards, the translation, and optional additional information such as gender, transliteration, script, and alternate forms. The parser extracts and retains all of this potentially valuable information.

Examples of translation templates:

[4] `http://cldr.unicode.org/`

[5] With the extra requirement that "there is clear notice in each modified Data File or in the Software as well as in the documentation associated with the Data File(s) or Software that the data or software has been modified."

[6] We later learned that McCrae et al. (2012) made a release of Wiktionary in the lemon format (`http://datahub.io/en/dataset/dbnary`). They did not, however, release the code they used to parse Wiktionary.

[7] Available from the Open Multilingual Wordnet Page: `http://casta-net.jp/~kuribayashi/multi/`.

[8] `http://dumps.wikimedia.org/`

- Finnish: {{t+|fi|sanakirja}}
- French: {{t+|fr|dictionnaire|m}}

To enable later processing, it is necessary to tie synonyms and translations to their corresponding short gloss via a unique key. Most parsers simply use an automatically generated surrogate key or a key based on the ordered position of data within a Wiktionary article. Since Wiktionary is constantly changing, the side effect of this approach is that data extracted from a specific snapshot of the Wiktionary database can only be meaningfully used in connection with other data extracted by the same parser from the exact same snapshot. To overcome this, we use a unique key that can be recreated from the data itself, which we call the **defkey**. To generate this key, we concatenate the language code, headword, part of speech, and the short gloss and use the `sha1` hash function (NIST, 2012) to create a unique 40-character hexadecimal string from the resulting text.

These defkeys are time and technology independent, so they allow the ability for researchers to efficiently share and compare results. Once a link is established between this defkey and a particular synset, translations added to Wiktionary at a later data can be automatically integrated into our multilingual wordnet. Conversely, if a Wiktionary contributor changes a short gloss, historical data connected to the old defkey is preserved while new data imported at a later time will not be incorrectly linked to an older definition.

Another feature of our parser is a feedback mode, which generates a report about poorly formated data that was encountered. These automatically generated reports can be used to create a quality-enhancing feedback loop with Wiktionary.

3.2.2 Linking Senses

Meyer and Gurevych (2011) showed that automatic alignments between Wiktionary senses and PWN can be established with reasonable accuracy and recall by combining multiple text similarity scores to compare a bag of words based on several pieces of information linked to a WordNet sense with another bag of words obtained from a Wiktionary entry. In our study we evaluated the potential for aligning senses based on common translations in combination with monolingual similarity features.

In this study we used 20 of the wordnets described in Section 2,[9] and the Wiktionary data obtained using the parser described in Section 3.2.1. Before searching for translation matches, we normalized the data to ensure the most accurate possible overlap count. First, article headwords were included as English translations of Wiktionary senses (along with synonyms). Then differences in language codes were rectified and translations containing symbolic characters or a mixture of roman and non-roman characters were marked to be ignored, save a few exceptions. This left approximately 1.4 million sense translations in 20 languages in our wordnet grid, and nearly 1.3 million Wiktionary translations in over 1,000 languages.

We then created a list of all possible alignments where at least one translation of a wordnet sense matched a translation of a Wiktionary sense. This represented a small percentage of the possible alignments, because definitions in Wiktionary that do not contain any translations were ignored in our study. Of more than 500,000 English definitions in Wiktionary, only about 130,000 presently have associated translations. The resulting graph contained over 700,000 possible sense alignments.

We calculated a number of similarity scores, the first two based on similarity in the number of lemmas, calculated using the Jaccard index:

$$\mathrm{sim_e}(s_n, s_k) = \frac{|E(s_k) \cap E(s_n)|}{|E(s_k) \cup E(s_n)|} \quad (1)$$

$$\mathrm{sim_a}(s_n, s_k) = \frac{|L(s_k) \cap L(s_n)|}{|L(s_k) \cup L(s_n)|} \quad (2)$$

Where s_k, s_n are concepts in Wiktionary and wordnet respectively,[10] $E(s)$ is the set of English lemmas for sense s and $L(s)$ is the set of lemmas in all languages.

As an initial pruning, we kept only matches where either: $\mathrm{sim_a} \geq 0.7$ **or** ($\mathrm{sim_e} \geq 0.5$ **and** $\mathrm{sim_a} \geq 0.5$) **or**, if ($|L(s_k) \cap L(s_n)| > 5$) then ($\mathrm{sim_e} \geq 0.5$ **and** $\mathrm{sim_a} \geq 0.45$). After applying these filters, approximately 220,000 alignment candidates remained.

We reviewed a random sample of 551 alignment candidates. Of these 136 were deemed correctly aligned. Another 48 we considered possibly close enough to produce valid translations for wordnet. All others were marked as incorrect alignments.

[9] We didn't use Chinese or Polish, as the wordnets were added after we had started the evaluation.

[10] Precisely, synsets in wordnet and senses in Wiktionary.

This development dataset was used to tune refined similarity scores.

$$\text{sim}_\text{t}(s_n, s_k) = \frac{|L(s_k) \cap L(s_n)|}{\sqrt{\alpha |L(s_k) \cup L(s_n)|}} \quad (3)$$

$$\text{sim}_\text{d}(s_n, s_k) = \frac{BoW(\texttt{wndef}) \cdot BoW(\texttt{wkdef})}{\|BoW(\texttt{wndef})\| \|BoW(\texttt{wkdef})\|} \quad (4)$$

$$\text{sim}_\text{c}(s_n, s_k) = \text{sim}_\text{t} + \beta \, \text{sim}_\text{c} \quad (5)$$

sim_t gives higher weight to concepts that link through more lemmas, not just a higher proportion of lemmas.

sim_d measures the similarity of the definitions in the two resources, using a cosine similarity score. We initially used the WordNet gloss and example sentence(s) for `wndef` and the short gloss from Wiktionary for `wkdef`. This improved the accuracy of the combined ranking score (sim_c), but since many of the short glosses are only one or two words, the sparse input often produced a sim_d score of zero even when the candidate alignment was correct. To improve the accuracy of the sim_d component, we also added in the long definitions.

Short glosses were aligned with long definitions using a similar approach to McCrae et al. (2012). First we search for a match where the short gloss was a substring of the full definition. If that failed to produce a single possible alignment, we aligned the short gloss with the full definition that produced the greatest cosine similarity score. Finally, where the short definition was blank and only a long definition was present, we aligned the two. The results of this alignment were less than 90% accurate, so to offset the effects of this noise we included both the full definition and the short gloss in `wkdef`. For `wndef` we used the WordNet gloss, example sentence(s), and synonyms. Even though the linking of definitions within Wiktionary left much to be desired, the increased amount of text improved the accuracy of the definition based similarity component of our ranking score.

Our combined ranking score (sim_c), based on both overlapping translations and a monolingual lexical similarity score, was able to outperform ranking based on either component in isolation. We expect that an improved alignment of short glosses to full definitions together with more accurate measures of lexical similarity such as described by Meyer and Gurevych (2011) would further improve the accuracy of a combined ranking score. We employed our combined ranking score first as a filter, where $\text{sim}_\text{c} \geq \tau_c$. The ranking score is then used to select the best match among competing alignments. Alignments are based on the belief that a definition within Wiktionary should only map to a single WordNet synset (if any at all). In theory, each WordNet synset should represent a meaning distinguishable from all other synsets. Because Wiktionary is organized according to lemma first, and sense second, multiple definitions in separate articles often map to the same synset. For example ***mortal*** "A human; someone susceptible to death", ***individual*** "A person considered alone ...", and ***person*** "A single human being; an individual" all align with $\text{someone}_{\text{n:1}}$ (00007846-n). However, two distinct definitions within the same Wiktionary entry should not map to the same WordNet sense. When there are multiple possible alignments where only one can be valid, sim_c is used to determine the best match.

In addition to using the combined ranking score as a filter, we found that we could obtain a small additional increase in accuracy without reducing recall by also requiring $\text{sim}_\text{t} \geq \tau_t$ **or** $\text{sim}_\text{d} \geq \tau_d$.

To determine ideal values for the weights and thresholds, we performed several grid searches. The parameters are interdependent and can produce reasonable results at a variety of points. Ideal values also depend on whether we wish to maximize accuracy or recall. α is set at 3.2 in order to achieve an ideal target threshold of $\tau_t = 1$. We finally chose values of $\beta = 0.7$ and $\tau_c = 0.71$ which gave a reasonable balance between accuracy and recall.

4 Results and Evaluation

We give the data for the 26 wordnets with more than 10,000 synsets in Table 2. There are a further 57 with more than 1,000; 133 with more than 100, 200 with more than 10 and 645 with more than 1 (although most of the very small languages appear to be simple errors in the language code entered into Wiktionary). Individual totals are shown for synsets and senses from the original wordnets, the data extracted from Wiktionary, and the merged data of the wordnets, Wiktionary and CLDR. We do not show the CLDR data in the table as it is so small, generally 500-600 synsets for the top languages. Overall there are 2,040,805 senses for 117,659 concepts, using over 1,400,000 words in over 1,000 languages.

The smaller wordnets are not of much practical use, but can still serve as the core of new

		Projects			Wiktionary			Merged (+CLDR)		
ISO	**Language**	**Synsets**	**Senses**	**Core**	**Synsets**	**Senses**	**Core**	**Synsets**	**Senses**	**Core**
eng	English	117,659	206,978	100	35,400	49,951	75	117,661	213,538	100
fin	Finnish	116,763	189,227	100	21,516	31,154	65	116,830	199,435	100
tha	Thai	73,350	95,517	81	2,560	3,193	17	73,595	97,390	81
fra	French	59,091	102,671	92	20,449	27,150	63	61,258	109,643	95
jpn	Japanese	57,179	158,064	95	12,685	19,479	52	59,112	166,617	96
ind	Indonesian	52,006	142,488	99	2,390	2,810	17	52,154	143,755	99
cat	Catalan	45,826	70,622	81	8,626	10,251	36	48,007	74,806	84
spa	Spanish	38,512	57,764	76	18,281	25,310	60	47,737	74,848	86
por	Portuguese	41,810	68,285	79	12,331	16,178	53	43,870	74,151	84
zsm	Standard Malay	42,766	119,152	99	2,833	3,744	19	43,079	120,686	99
ita	Italian	34,728	60,561	83	14,605	18,710	53	38,938	68,827	87
eus	Basque	29,413	48,934	71	1,693	1,943	11	29,965	49,945	72
pol	Polish	14,008	21,001	30	10,888	13,431	46	20,975	30,943	55
glg	Galician	19,312	27,138	36	2,492	2,871	15	20,772	29,136	42
fas	Persian	17,759	30,461	41	4,229	5,443	26	20,766	35,318	55
rus	Russian	0	0	0	19,983	33,716	64	20,138	34,009	64
deu	German	0	0	0	19,675	29,616	64	19,857	29,884	64
cmn	Mandarin Chinese	4,913	8,069	28	12,130	19,079	49	15,490	27,113	60
arb	Standard Arabic	10,165	21,751	48	6,892	9,337	38	14,861	31,337	63
nld	Dutch	0	0	0	13,741	19,709	56	13,950	20,003	56
ces	Czech	0	0	0	12,802	15,493	54	13,030	15,813	54
swe	Swedish	0	0	0	12,000	16,226	51	12,221	16,512	51
ell	Modern Greek	0	0	0	10,308	13,071	44	10,549	13,472	44
dan	Danish	4,476	5,859	81	7,290	8,931	35	10,328	13,551	85
nob	Norwegian Bokmål	4,455	5,586	79	7,262	9,170	35	10,322	13,612	83
hun	Hungarian	0	0	0	9,964	12,699	45	10,213	13,029	45

Core shows the percentage coverage of the 5,000 core concepts.

Table 2: Merged Wordnets (with more than 10,000 entries)

projects. The bigger wordnets show the data from Wiktionary (and to a lesser extent CLDR) having only a small increase in the number of senses. The biggest change is for the medium size projects, such as Persian or Arabic, which end up with much better coverage of the most frequent core concepts. Major languages such as German or Russian, which currently do not have open wordnets get good coverage as well.

The size of the mapping table is the same as the number of English senses linked (49,951 senses). We evaluated a random sample of 160 alignments and found the accuracy to be 90% (Wiktionary sense maps to the best possible wordnet sense).

We then evaluated samples of the wordnet created from Wiktionary for several languages. For each language we choose 100 random senses, then checked them against existing wordnets.[11] For all unmatched entries, we then had them checked by native speakers. The results are given in Table 3. The sense accuracy is higher than the mapping accuracy: in general, entries with more translations are linked more accurately, thus raising the average precision. During the extraction and evaluation, we noticed several language specific features: for example, Serbo-Croatian had a mixture of Cyrillic and Latin entries. For languages where one script was clearly dominant, we kept only that, but really these decisions should be done for each language by a native speaker.

Language	**% Matched**	**% Good**
Chinese*	46	97
Serbo-Croation*,**	0	91
Czech*	0	99
English	89	92
German*	19	85
Indonesian	69	97
Korean*	0	96
Japanese	56	90
Russian*	0	99
Average		94.0

Table 3: Precision of Wiktionary-based Wordnets

* Not used to build the mapping from wordnet to Wiktionary.
** We allow terms used in either Serbian or Croatian.

We make the data available in two ways. The first is a set of downloads. Each language has up to three files: the data from the wordnet project (if it exists), the data from the CLDR and the data from Wiktionary. They are kept separate in order

[11] For Chinese we use the wordnet from Xu et al. (2008), which is free for research but cannot be redistributed. For German we used Euro WordNet (Vossen, 1998).

to keep the licenses as free as possible. The second is as two on-line searches: one using only the data from the projects, and one with all the data combined. The combination is done by simple union.[12] We maintain this separation as we cannot guarantee the quality of the automatically extracted data. Because the raw data is there it is possible to combine them in other ways. The simple structure is easy to manipulate, and there is code to use this style of data with the popular tool kit NLTK (Bird et al., 2010).

5 Discussion and Future Work

We have created a large open wordnet of high quality (85%–99% measured on senses). Twenty six languages have more than 10,000 concepts covered, with 42–100% coverage of the most common core concepts. The data is easily downloadable with minimal restrictions. The overall accuracy is estimated at over 94%, as most of the original wordnets are hand verified (and so should be 100% accurate). The high accuracy is largely thanks to the disambiguating power of the multiple translations, made possible by the many open wordnets we have access to.

Because we link senses between wordnet and Wiktionary and then use the translations of the sense, manually validating this mapping will improve the entries in multiple languages simultaneously. As the Wiktionary-wordnet alignment mapping is linked to persistent keys it will remain useful even as the resources change. Further, it can be used to identify and add missing senses to wordnet: unmapped Wiktionary entries are candidates for new concepts.

The Universal Wordnet (UWN: de Melo and Weikum, 2009) brings in data from even more resources, and combines them to make a larger resource, choosing parameters with slightly lower precision (just under 90%). It is further linked to Wikipedia, adding many named entities. We expect that our work is complementary. Because we use a different approach, it would be possible to merge the two if the licenses allowed us to. However, since the CC BY-SA and CC-BY-NC-SA licenses are mutually exclusive, the two works cannot be combined and rereleased unless relevant parties can relicense the works. There is no easy way to improve UWN beyond checking each and every entry, which is expensive. An advantage of our approach, noted above, is that we can validate the sense matches for English and the accuracy percolates down to all the languages.

Integrating data from the most recent version of Wiktionary can be done simply and takes a few hours. It is therefore feasible to update the downloadable data regularly. Improvements in either the wordnet projects or Wiktionary (or both) can also result in improved mappings. We further hope to take advantage of ongoing initiatives in the global wordnet grid to add new concepts not in the Princeton WordNet, so that we can expand beyond an English-centered world view.

By making the data from multiple sources easily available with minimal restrictions, we hope that it will be easier to do research that exploits lexical semantics. In particular, we make the data easily accessible to the original wordnet projects, some of whom have already started to merge it into their own resources. We cannot check the accuracy of data in all languages, nor, for example, check that synsets have the most appropriate lemmas associated with them. Many languages have their own orthographic issues (for example a choice of scripts, or the choice to include vowels or not). Our automatic extraction does not deal with these issues at all. This kind of language specific quality control is best done by the individual wordnet projects.

We also consider it important to keep feeding data back to the individual wordnet projects, as much of the innovative research comes from them: the class/instance distinction from PWN; the distinction between rigid and non-rigid synsets from the Kyoto Project; domain mappings from the MultiWordNet (Pianta et al., 2002); representing orthographic variation from the Japanese Wordnet (Kuroda et al., 2011); combining close languages from the Wordnet Bahasa (Nurril Hirfana et al., 2011); and so on. For all of these reasons, we do not consider automatic extraction from/linking to Wiktionary a substitute for building languages specific wordnets.

Further work that this data should allow us to do include: automatically producing a list of bad data found in Wiktionary that can be used by Wiktionary editors to correct errors; and finding gaps in wordnet by identifying senses in Wiktionary that have a large number of translations, but fail to have any significant alignment with existing wordnet synsets.

[12] `http://casta-net.jp/~kuribayashi/multi/`

We currently only link through the English Wiktionary and its translations. It should be possible to expand the multilingual wordnet in the same way using Wiktionaries in other languages, which we would expect to improve coverage.

Finally, Wiktionary contains a lot of useful information we are not currently using (information on gender, transliterations, pronunciations, alternative spellings and so forth). We can also think of the aligned definitions as a paraphrase corpus for English.

We have devoted more space than is usual for a computational linguistics paper to issues of licensing and sustainability. This is deliberate: we feel papers about lexical resources should be clear about licensing, and that it should be considered early on when creating new resources. There are strong arguments that open data leads to better science (Pederson, 2008), and it has been shown that open resources are cited more (Bond and Paik, 2012). In addition, how to maintain resources over time is a major unsolved problem. We consider it important that our wordnet is not just large and accurate but also maintainable and as accessible as possible.

6 Conclusions

We have created an open multilingual wordnet with over 26 languages. It is made by combining wordnets with open licences, data from the Unicode Common Locale Data Repository and Wiktionary. Overall there are over 2 million senses for 117,659 concepts, using over 1.4 million words in hundreds of languages.

Acknowledgments

We would like to thank the following for their help with the evaluation: Le Tuan Anh, František Kratochvíl, Kyonghee Paik, Zina Pozen, Melanie Siegel, Stefanie Stadler, Bilyana Shuman, Liling Tan and Muhammad Zulhelmy bin Mohd Rosman.

References

Stephen Bird, Ewan Klein, and Edward Loper. 2010. *Nyumon Shizen Gengo Shori [Introduction to Natural Language Processing]*. O'Reilly. (translated by Hagiwara, Nakamura and Mizuno).

W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, M. Bertran, and C. Fellbaum. 2006. The Arabic wordnet project. In *Proceedings of LREC 2006*.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64–71.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to WordNet. In *Proceedings of the Third Global WordNet Meeting*. Jeju.

Thatsanee Charoenporn, Virach Sornlerlamvanich, Chumpol Mokarat, and Hitoshi Isahara. 2008. Semi-automatic compilation of Asian WordNet. In *14th Annual Meeting of the Association for Natural Language Processing*, pages 1041–1044. Tokyo.

Jordi Daude, Lluis Padro, and German Rigau. 2003. Validation and tuning of Wordnet mapping techniques. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*. Borovets, Bulgaria.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522. ACM, New York, NY, USA.

Gerard de Melo and Gerhard Weikum. 2010. Towards universal multilingual knowledge bases. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global WordNet Conference (GWC 2010)*, pages 149–156. Narosa Publishing, New Delhi, India.

Valeria de Paiva and Alexandre Rademaker. 2012. Revisiting a Brazilian wordnet. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.

Christiane Fellbaum and Piek Vossen. 2012. Challenges for a multilingual wordnet. *Language Resources and Evaluation*, 46(2):313–326. Doi=10.1007/s10579-012-9186-z.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repos-

itory version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.

Valérie Hanoka and Benoît Sagot. 2012. Wordnet creation and extension made simple: A multilingual lexicon-based approach using wiki resources. In *Proceedings of LREC 2012*. Istanbul.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing*, 24(2):14–23. (in Chinese).

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Toru Ishida. 2006. Language grid: An infrastructure for intercultural collaboration. In *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pages 96–100. URL `http://langrid.nict.go.jp/file/langrid20060211.pdf`, (keynote address).

Kow Kuroda, Takayuki Kuribayashi, Francis Bond, Kyoko Kanzaki, and Hitoshi Isahara. 2011. Orthographic variants and multilingual sense tagging with the Japanese WordNet. In *17th Annual Meeting of the Association for Natural Language Processing*, pages A4–1. Toyohashi.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet — wordnet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17:119–140. In Swedish with an English abstract.

John McCrae, Philipp Cimiano, and Elena Montiel-Ponsoda. 2012. Integrating wordnet and wiktionary with lemon. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellman, editors, *Linked Data in Linguistics*. Springer.

Christian M. Meyer and Iryna Gurevych. 2011. What psycholinguists know about chemistry: Aligning wiktionary and wordnet for increased domain coverage. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 883–892.

Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 258–267. Singapore.

Mortaza Montazery and Heshaam Faili. 2010. Automatic Persian wordnet construction. In *23rd International conference on computational linguistics*, pages 846–850.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

NIST. 2012. Secure hash standard (shs). Fips pub 180-4, National Institute of Standards and Technology.

Noam Ordan and Shuly Wintner. 2007. Hebrew wordnet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.

B.S Pedersen, S. Nimb, J. Asmussen, N. Sørensen, L. Trap-Jensen, and H. Lorentzen. 2009. DanNet — the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*.

Ted Pederson. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *In Proceedings of the First International Conference on Global WordNet*, pages 293–302. Mysore, India.

Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wroclaw University of Technology Press. URL `http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf`, (ISBN 978-83-7493-476-3).

Chiew Kin Quah, Francis Bond, and Takefumi Yamazaki. 2001. Design and construction of a machine-tractable Malay-English lexicon.

In *Asialex 2001 Proceedings*, pages 200–205. Seoul.

Ervin Ruci. 2008. On the current state of Albanet and related applications. Technical report, University of Vlora. (`http://fjalnet.com/technicalreportalbanet.pdf`).

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich, and Hitoshi Isahara. 2009. Thai wordnet construction. In *Proceedings of The 7th Workshop on Asian Language Resources (ALR7), Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP)*,. Suntec, Singapore.

Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology*, 7(1–2):9–34.

Unicode, Inc. 2012. Unicode, Inc. license agreement - data files and software. `http://www.unicode.org/copyright.html`.

Piek Vossen, editor. 1998. *Euro WordNet*. Kluwer.

Piek Vossen. 2005. Building wordnets. `http://www.globalwordnet.org/gwa/BuildingWordnets.ppt`.

Wikimedia. 2013. List of wiktionaries. `http://meta.wikimedia.org/w/index.php?title=Wiktionary&oldid=4729333`. (accessed on 2013-02-14).

Wikipedia. 2013. Wikipedia — wikipedia, the free encyclopedia. URL `http://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=552515903`, [Online; accessed 30-April-2013].

Renjie Xu, Zhiqiang Gao, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual Chinese-English WordNet. In *3rd Asian Semantic Web Conference (ASWC 2008)*, pages 302–341.