English WordNet:
A new open-source Wordnet for English

John P. McCrae, Ewa Rudnicka and Francis Bond

Introduction

Wordnets have become one of the most popular dictionaries for use in natural language processing (NLP) and other areas of language technologies. This is primarily due to their structure as a graph of words, that is much easier for computers to understand than the traditional form of a dictionary. The first wordnet was introduced by Arthur Miller (Miller 1995) and later extended by Christiane Fellbaum (Fellbaum 1998) at Princeton University before finally being released in its definitive form as Princeton WordNet 3.0 in 2006. However, since then there has only been a single maintenance release of the resource (3.1) in 2011, that actually reduced the number of words it covered. Meanwhile, interest and use of wordnets have grown with many projects around the world creating new wordnets for languages other than English as well as projects adding extensions to Princeton WordNet such as extending it with sentiment information (Esuli and Sebastiani 2006), encyclopedic information (Navigli and Ponzetto 2012) and pronouns and exclamatives (Da Costa and Bond 2016), and providing domain-specific terminology (McCrae, Wood, and Hicks 2017). Furthermore, it is clear that the English language has changed in the last 14 years and Princeton WordNet does not cover recent neologisms and other language usage changes, which are important for many of the social media analytics tasks that we wish to apply wordnets to. Moreover, perhaps one of the biggest criticisms of Princeton WordNet has been that it contains many errors (McCrae and Prangnawarat 2016) and has, at times, overly fine or coarse sense distinctions (Hovy et al. 2006).

Given the lack of change in Princeton WordNet, in spite of the abundant criticisms, we decided to make a 'fork' of the Princeton WordNet, to create a new open-source project called English WordNet (EWN). This project aims to produce the highest quality and most complete wordnet for English and to do so in an open manner. This is implemented by means of a GitHub repository with a collection of XML files that are clear and can easily be edited by anyone. The project accepts suggestions from any parties and so far has been very active with over 650 commits over 500 issues over the course of two years. This has lead to over 18,500 individual improvements over the

John P. McCrae is a research lecturer at the Data Science Institute at the National University of Ireland Galway and a member of the SFI Insight Research Centre for Data Analytics. He holds a PhD from SOKENDAI University (National Institute of Informatics, Tokyo). He is the coordinator of the H2020 project Prêt-à-LLOD on linguistic linked open data and leads the task on linked data in the ELEXIS infrastructure H2020 project, and holds an IRC (Irish Research Council) consolidator laureate award on NLP for minority and historical languages and is a board member of the Global WordNet Association.
john.mccrae@insight-centre.org

Princeton WordNet, producing a resource that is clearly of much better quality and more comprehensive than the previous releases that have been available to date.

In this article, we provide a brief description of the idea of wordnets and how they are frequently used in natural language processing for readers who may not be familiar with this form of dictionary. Then, we describe the development methodology we have for this dictionary and how we have built and adapted to the growing community of English WordNet users. We then describe the resource of English WordNet and the changes over Princeton WordNet in its two releases so far. Finally, we detail our future plans for this wordnet and make some concluding remarks.

Wordnets are a form of dictionary that aim to make information more easily processable for computers. The primary unit of a wordnet is a set of synonyms or a *synset*, consisting of a list of words that in some context can be substituted for each other. These synsets then form the nodes of a graph, which is connected by edges, consisting of relations such as *hypernym*, indicating a broader/narrower relation, *antonym*, indicating opposition, and *meronym*, indicating a part/whole relation. A word may be part of multiple synsets and as such, we refer to the word within a given synset as a *sense* of the word. An example of such a graph is shown in Figure 1.

Princeton WordNet and most other wordnets cover only four parts-of-speech: noun, verb, adjective and adverb. The nouns are grouped into a hierarchy, where every term is ultimately a hyponym of a single word 'entity'. Verbs similarly are grouped into hierarchies, however, there is no overall supreme concept for verbs and the graph is more disconnected. For adjectives, the structure is generally based around a 'dumbbell' model, where adjectives are grouped into pairs of antonyms, such as 'hot'-'cold', and then 'satellite' adjectives that are related to the meaning of these adjectives, such as 'scorching' or 'frosty', are connected to one end of the dumbbell with a *similar* relation. Alternatively, adjectives may be classified as *pertainyms*, whose meaning is defined by 'of or relating to' a noun, such as 'French' to 'France'. For adverbs, there is little structure and many adverb synsets have no connections in the graph.

The graph-based nature of wordnets has made them highly amenable to NLP applications and a number of methods have been developed that exploit this. For example, word similarity can be computed by simply calculating how many edges must be followed to connect two words (Wu and Palmer 1994) and more sophisticated methods have been built on this principle (Lin and Sandkuhl 2008). Moreover,

Ewa Rudnicka is a research associate at the Department of Computational Intelligence, Wroclaw University of Science and Technology, Poland, and a member of CLARIN-PL Language Technology Centre. She holds a PhD and an MA in comparative linguistics from the University of Wroclaw (Faculty of Languages, Department of English). She is the coordinator of a team of lexicographers working on the mapping between plWordNet and Princeton WordNet, and building an extension to the latter, enWordNet, and a member of the Global Wordnet Association. Her research interests cover lexicography, semantics, theory of equivalence, and natural language processing.
eva.rudnicka@pwr.edu.pl

Princeton WordNet is still the most widely used resource for *word sense disambiguation*, the task of deciding which sense of a word is used in a given context, and wordnets are still the basis of most evaluations in this area (Navigli, Jurgens and Vannella 2013). Even with the recent developments in the field of NLP, relating to the use of neural networks and other methods, there has been interest in exploiting the graph structure of wordnets to develop neural networks (Kutuzov et al. 2018) and embeddings (Rothe and Schütze 2015).

Open-source methodology of EWN

English WordNet has adopted an open-source methodology for the development of the wordnet, meaning that anyone can comment and suggest changes, although these changes are implemented by a core team of developers. There are principally two ways to contribute to EWN, either directly by suggesting changes to the XML through a method called a *pull request*, that is a standard part of open-source development, or by making an *issue*, that is a report of a bug. We have found that the vast majority of suggested changes are made by opening an issue. These suggestions are then categorized by the type of change that is requested, for example adding/removing a relation, updating a definition or example of usage, adding, removing, merging or splitting a synset, or another technical issue. We find that most issues refer to merging synsets, perhaps because wordnet tends to split word senses too much, but can often be resolved by making the definitions more distinct. In addition, there have been many requests for new synsets to be added and to accommodate this we have developed guidelines that determine when a new term should be added to the wordnet.

- Concepts should be significant and represent general English usage. English WordNet does not need to include the name of every place, person and organization in the world. Such things are better handled by other projects, such as Wikidata.
- Terms should not be compositional, that is the meaning of a multiword expression cannot be inferred from its words, or a single word is not derived by the obvious use of a prefix or suffix.
- The word (or sense) should be distinct from other synsets already in the wordnet.
- It should be possible to give a clear textual definition of the concept and to link it to at least one other concept already in the wordnet.

Francis Bond is an Associate Professor in Linguistics and Multilingual Studies at Nanyang Technological University, Singapore and the co-coordinator of NTU's Digital Humanities Cluster. He holds a BA, BEng and PhD from the University of Queensland. He is an active member of the Deep Linguistic Processing with HPSG Initiative (DELPH-IN) and the Global WordNet Association, has developed and released wordnets for Chinese, Japanese, Malay and Indonesian, and coordinates the Open Multilingual Wordnet. His main research interest is in natural language understanding.
bond@ieee.org

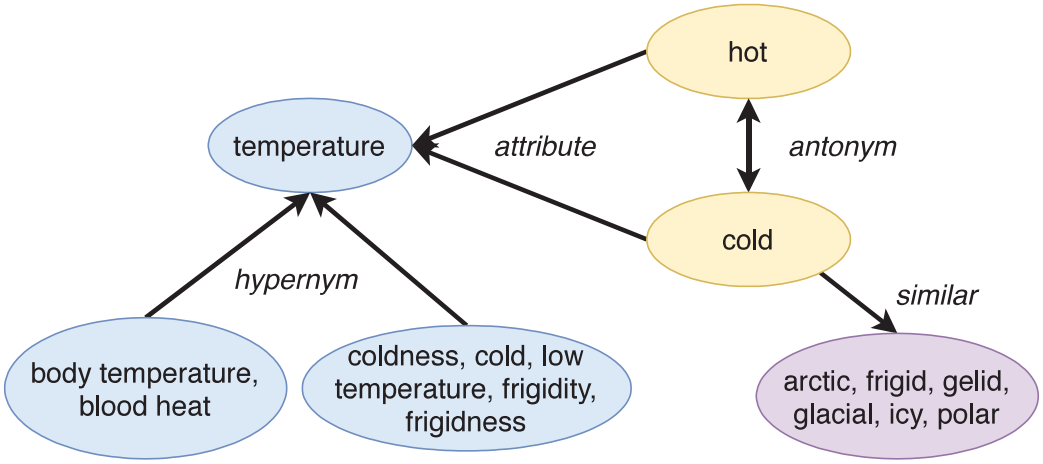

Figure 1. An example of a wordnet graph, showing 'temperature' and its hypernyms, the dumbbell of 'hot' and 'cold' and a satellite adjective

- In difficult cases, we look for clear distinctions in the hypernym to distinguish similar concepts, such as for 'wood' by consistently distinguishing between a tree (an organism) and its wood (a material) or by finding collocations that clearly distinguish this sense.

A key goal is to ensure that there is backwards compatibility between these releases of EWN and the previous Princeton releases. We achieve this by also releasing the data in the form of the database files that are used by the Princeton WordNet tools. This can create some issues in that this format uses the *offset*, that is the number of bytes in the file that need to be read to reach the start of an entry, to identify synsets. For English WordNet, we have fixed the identifiers to be the offsets of the Princeton WordNet 3.1 release and in fact, use random numbers for new synsets so it would be impossible and impractical to keep these in-sync with the release. Otherwise, we try to keep all of the features of Princeton WordNet's structure as is, even if some aspects may be unnecessary, complex or scientifically questionable.

To date, there have been two releases of English WordNet, the 2019 and 2020 edition. These have expanded the scope of the project and

	Princeton 3.1	EWN 2019	EWN 2020
Synsets	117,791	117,791	120,054
Lemmas	159,015	159,789	163,079
Senses	207,272	208,353	211,864
Relations	378,203	378,201	383,825

Table 1. Size and coverage of Princeton WordNet 3.1 and the two releases of English WordNet

while they have obviously introduced many new changes, the focus of the work has been on improving the quality of the resource. In fact, we found nearly 2,000 typos in the text of PWN, and even in one case a misspelt lemma!

Another major source of changes was the inclusion of external data from other sources. We directly included other English wordnets, including enWordNet developed as part of plWordNet (Rudnicka, Witkowski and Kaliński 2015) and Colloquial WordNet (McCrae, Wood and Hicks 2017), with some modification to better fit the structure of this wordnet. Secondly, we incorporated updated definitions from the Open Multilingual WordNet project (Bond and Foster 2013) and also used the linking to Wikipedia (McCrae 2018) to add extra lemmas for many concepts. Finally, we fixed many minor errors related to issues such as examples which do not use any lemma in the synset.

Future plans

English WordNet is an expanding project and we intend to continue to develop the resource through the open-source methodology. There have been several areas that have been identified as key long-term areas to improve the resources. Firstly, the modelling of adjectives and adverbs is, as discussed above, quite unusual and adjectives and adverbs have far fewer links and more disconnected nodes in the graph than for nouns or verbs. Adopting a new structure would be much more preferable (Mendes 2006), and finding similar ways to define adverbs and their relations to the noun and verb hierarchies would enhance usability for NLP applications that depend on these links. Secondly, we are working to improve the methodology for developing the wordnet, in particular, there has been much discussion in the community about moving on from the XML model to something less verbose and more readable and the use of YAML markup is likely to be adopted. As an example we compare the current XML markup with the proposed YAML form, which significantly reduces the size of the file.

In addition, we have a browsing interface available at https://en-word.net/ which provides a searchable interface to the most recent interface and provides a linked data version of the data in RDF using the OntoLex-Lemon model (Cimiano, McCrae and Buitelaar 2016). As such, the YAML format is intended to be an internal working format with releases still made according to the standards such as the LMF XML format, and OntoLex-Lemon. An example of the data encoding is available in Figure 2.

```
<LexicalEntry id="ewn-dictionary-n">
 <Lemma writtenForm="dictionary"
    partOfSpeech="n"/>
 <Sense id="ewn-dictionary-n-06430544-01"
    n="0" synset="ewn-06430544-n"
    dc:identifier="dictionary%1:10:00::"/>
</LexicalEntry>
<Synset id="ewn-06430544-n" ili="i70226"
   partOfSpeech="n"
   dc:subject="noun.communication">
 <Definition>a reference book containing an alphabetical list of
words with information about them</Definition>
 <SynsetRelation relType="hypernym"
       target="ewn-06430336-n"/>
 <SynsetRelation relType="mero_part"
       target="ewn-06311813-n"/>
</Synset>
```

```
06430544-n:
 definitions:
 - a reference book containing an
  alphabetical list of words
  with information about them
 entries:
 - dictionary%1_10_00
 - lexicon%1_10_00
 hypernym:
 - 06430336-n
 ili: i70226
 mero_part:
 - 06311813-n
 pos: n
```

Figure 2. An example of the XML and YAML encoding of the English WordNet data as available on Github

Furthermore, tools for supporting changes in English WordNet and validating the consistency are already deployed and continue to be developed. Finally, we would like to move on from the model of a single monolithic dictionary and support a network of wordnets, including domain-specific wordnets or large-scale encyclopedic resources that could be of use to a wide range of tasks, although this would create further issues with maintaining and integrating such a wide range of tasks.

Conclusion

English WordNet is an open-source fork of the Princeton WordNet, whose aim is principally to ensure that there is an English wordnet which is up-to-date and can be of the highest quality, as the many users of wordnets can easily contribute changes and improvements back to the project. We have done this in a simple way, by providing a GitHub repository for simple XML documents. This has proven successful with over 18,500 changes and many contributions from all sides. We plan to continue to develop this resource and hope that it continues to be one of the core dictionaries for NLP applications. Further, while this project is intended to be limited to the English language we hope that this methodology can be adopted by wordnets for other languages and support linking and connecting to create multilingual resources such as through the Open Multilingual WordNet.

References

Bond, Francis and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1352–62. Sofia, Bulgaria: Association for Computational Linguistics.

Cimiano, Philipp, John P. McCrae and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community Report. W3C. https://www.w3.org/2016/05/ontolex/.

Da Costa, Luis Morgado and Francis Bond. 2016. Wow! What a Useful Extension! Introducing Non-Referential Concepts to WordNet. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4323–28.

Esuli andrea and Fabrizio Sebastiani. 2006. Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining. In *LREC*, 6:417–22. Citeseer.

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

Hovy, Eduard, Mitch Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel. 2006. "OntoNotes: The 90% Solution." In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 57–60.

Kutuzov, Andrey, Mohammad Dorgham, Oleksiy Oliynyk, Chris Biemann and Alexander Panchenko. 2018. Learning Graph Embeddings from WordNet-Based Similarity Measures. *arXiv [cs. CL]*. arXiv. http://arxiv.org/abs/1808.05611.

Lin, Feiyu and Kurt Sandkuhl. 2008. A Survey of Exploiting WordNet in Ontology Matching. In *Artificial Intelligence in Theory and Practice II*, 341–50. Springer US.

McCrae, John P. 2018. Mapping WordNet Instances to Wikipedia. In *Proceedings of the 9th Global WordNet Conference*.

McCrae, John P. and Narumol Prangnawarat. 2016. Identifying Poorly-Defined Concepts in WordNet with Graph Metrics. In *Proceedings of the First Workshop on Knowledge Extraction and Knowledge Integration (KEKI-2016)*.

McCrae, John P., Ian Wood and Amanda Hicks. 2017. The Colloquial WordNet: Extending Princeton WordNet with Neologisms. In *Proceedings of the First Conference on Language, Data and Knowledge (LDK2017)*, 194–202.

Mendes, Sara. 2006. Adjectives in WordNet.PT. In *Proceedings of the Global WordNet Conference*, edited by Petr Sojka, Key-Sun Choi, Christiane Fellbaum and Piek Vossen, 225–30. Citeseer.

Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38 (11): 39–41.

Navigli, Roberto, David Jurgens and Daniele Vannella. 2013. Semeval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 222–31.

Navigli, Roberto and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence* 193 (Supplement C): 217–50.

Rothe, Sascha and Hinrich Schütze. 2015. AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1793–1803.

Rudnicka, Ewa Katarzyna, Wojciech Witkowski and Michał Kaliński. 2015. Towards the Methodology for Extending Princeton Wordnet. *Cognitive Studies| Études Cognitives*, no. 15: 335–51.

Wu, Zhibiao and Martha Palmer. 1994. Verbs Semantics and Lexical Selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, 133–38. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics.