The Colloquial WordNet: Extending Princeton WordNet with Neologisms

John P. McCrae[1], Ian Wood[1], and Amanda Hicks[2]

[1] Insight Center for Data Analytics, National University of Ireland, Galway
john@mccr.ae, ian.wood@insight-centre.org
[2] Department of Health Outcomes and Policy, University of Florida
aehicks@ufl.edu

Abstract. Princeton WordNet is one of the most important resources for natural language processing, but has not been updated for over ten years and is not suitable for analyzing the fast moving language as used on social media. We propose an extension to WordNet, with new terms that have been found from Twitter and Reddit, and cover language usage that is emergent or vulgar. In addition to our methodology for extraction, we analyze new terms to provide information about how new words are entering the English language. Finally, we discuss publishing this resource both as linguistic linked open data and as part of the Global WordNet Association's Interlingual Index.

Keywords: wordnet, neologisms, slang, linked data, lexicography

1 Introduction

Princeton WordNet (PWN) [9] is the most widely used lexical resource in natural language processing. However, it has not been updated significantly since the release of Version 3.0 in 2006. As such, there are many new terms that have entered the English language, which are not covered by this resource. Yet many applications, especially in sentiment analysis, base their analysis on texts extracted from social media platforms, where the use of language is often quite distinct from the general language that is covered by WordNet. Moreover, social media has allowed communities to gather around specific topics of interest [11] and often the language exhibits distinct features [10] and a vocabulary that is not captured by WordNet.

In this paper, we present the initial version of a new resource we call the Colloquial WordNet, which extends Princeton WordNet to work better in new domains, especially those such as internet forums and messaging services such as Twitter. Furthermore, we extend on some of the challenges in this domain and provide not only traditional lexical entries, but also lists of misspellings, abbreviations and common errors. Furthermore, we investigate the construction of neologisms in social media in comparison to language used in general and technical domains.

2 Methodology

2.1 Corpus Preparation

We extracted a corpus from two social media websites: Twitter, where text was gathered using the Twitter sample API endpoint[3] between February 2nd and 22nd 2016; and Reddit, where we extracted data from the top 1000 most popular forums ('subreddits') using a webpage crawler[4]. In total, we collected 255,908 Reddit posts (3.4 million tokens) and 3,018,180 Twitter posts (29.8 million tokens).

Our approach for selecting terms was based on the ratio of the frequency of terms in the Reddit or Twitter corpus relative to a background corpus, in particular, the *Google Web Trillion Word Corpus*[5]. To improve the ranking of this ratio, we discarded all terms that did not occur at least 10 times in the Reddit or Twitter corpus and set the frequency of terms not found in the background corpus or in the lowest decile to the highest value in the lowest decile. We then filtered terms to only those that occurred in Urban Dictionary[6], that occur in all lowercases more frequently than otherwise and other filters to remove simple non-terms (such as phrases starting with 'a' or 'the'). This gave us the ability to find terms that would be relevant with high precision, and our annotators accepted 61.3% of terms as worthy of inclusion in the lexicon, among the 500 highest scoring terms.

2.2 Annotation Procedure

Using the terms selected as potentially relevant, the annotators were asked to create entries using the interface shown in Figure 1. The first decision made by the annotator was the status of the term, which could be one of the following:

General A term that is generally used in the language and would be suitable for inclusion in PWN. This covers some new terms such as 'steampunk' or 'hoverboard' that cover novel concepts. A few times surprising gaps in PWN were found for example a sense of the verb 'pick' in 'lock picking'. This can mean that novel senses are added to existing Princeton WordNet entries.

Novel This is for terms that the annotators believed may not be stable in the language, in that they are extremely colloquial, e.g., 'bestie' (best friend) or they refer to a current cultural phenomenon, e.g., 'twerk' and 'dab' (popular dance moves). As such terms may not remain in the language for long they are tagged in the data as novel terms.

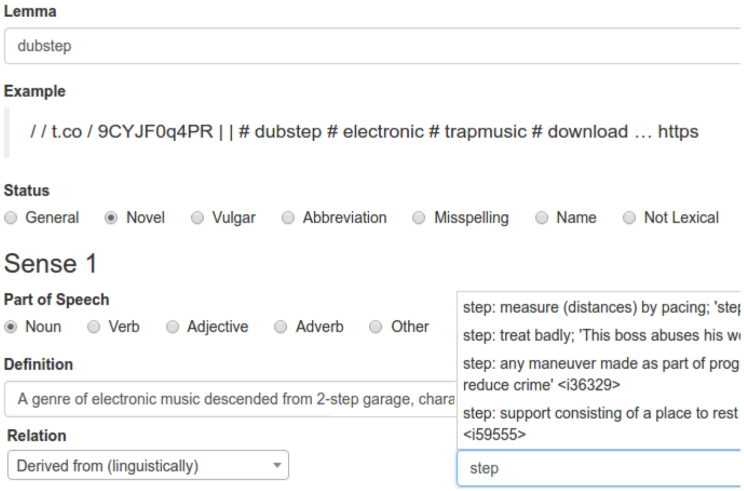
Vulgar This covers both terms that use vulgar language, refer to sexual acts or are defamatory (racist, sexist, etc.). A significant number of the tweets in our corpus were advertising pornography or sexual services, resulting in many vulgar terms in the output.

[3] This end point provides a sample of approximately 1% of all tweets.

[4] https://github.com/lucasdnd/simple-reddit-crawler

[5] Compiled at http://norvig.com/ngrams/

[6] http://www.urbandictionary.com



Lemma

dubstep

Example

/ / t.co / 9CYJF0q4PR | | # dubstep # electronic # trapmusic # download … https

Status

○ General ● Novel ○ Vulgar ○ Abbreviation ○ Misspelling ○ Name ○ Not Lexical

Sense 1

Part of Speech

● Noun ○ Verb ○ Adjective ○ Adverb ○ Other

Definition

A genre of electronic music descended from 2-step garage, chara

Relation

Derived from (linguistically)

step: measure (distances) by pacing; 'step

step: treat badly; 'This boss abuses his w

step: any maneuver made as part of prog reduce crime' <i36329>

step: support consisting of a place to rest <i59555>

step

Fig. 1. An example entry, 'dubstep', showing the status definition and one link being created

Abbreviation The term is an abbreviation.

Misspelling The term is a misspelled version of a standard English word or phrase.

Name The term is a name of a person, organization or place. Note we classed terms derived from names, e.g., 'belieber' (a fan of Justin Bieber) as novel words.

Not Idiomatic The algorithm detected some short expressions as new terms when they were in fact just collocations, e.g., 'can i get'.

Error This was used for technical errors, e.g., 'nbsp' (the HTML entity for non-breaking space).

The next step of the annotation involved either selecting an existing synset in PWN to which the word referred or writing a novel English definition for the term as well as deciding the part of speech. Note that following Morgado da Costa and Bond [6] we also allow the annotation of interjection expressions such as 'oh' or 'haha' as these are useful for understanding emotion and meaning in social media texts. If the word had multiple meanings the annotator could create multiple senses, each with their own definition and part-of-speech. The final step for the annotator was to add links from the new synset to any other synsets in PWN. This was supported by an interactive selection tool (see Figure 1) and all the standard relations could be selected. In addition, we included two new relations that were useful, firstly a 'loanword' as many of our neologisms were words from other languages and secondly an 'emotion' property to indicate what feeling is expressed by the meaning of a word.

	Size		Size
Entries	428	Non-entries	1340
- General	83	- Non-Idiomatic	392
- Novel	181	- Errors	83
- Vulgar	46	- Proper Nouns	336
- Interjections	117	- Abbreviations	184
Synsets	430	- Misspellings	345
Sense Relations	408		
Synset Relations	365		

Table 1. The size of the resource in terms of elements it contains

3 Results

3.1 Resource creation

The overall statistics for the resource are presented in Table 1, where we describe the number of new entries found, broken down into the categories (General, Novel, Vulgar) described above as well as the number of synsets and relations between synsets in the new graph. In addition, we provide the non-lexical items that we found during the construction of the resource, which represent one of the major forms of the elements found. While we see that very few of the forms are true errors, there are a large number of items in the categories of non-idiomatic phrases, misspellings and proper nouns. This is slightly surprising given that a lexicon (Urban Dictionary) was used to filter out terms that were not suitable for inclusion in the dictionary and this demonstrates the unreliability of Urban Dictionary as a base resource.

3.2 An analysis of Neologisms

In addition, we collected information during the annotation procedure about the origin of neologisms that have been defined. We did this by classifying neologisms into the following groups:

Novel Sense This describes a novel meaning that a word has acquired recently. For example the noun 'post' in the sense of a 'forum post' is a sense that does not match the existing senses of the noun in PWN 3.1[7]. In addition, we also count cases where a word has changed part-of-speech as new senses, such as the verb 'favorite', which is only listed as an adjective in PWN 3.1.

Multiword Expression The standard method of constructing new terms is the combination of two or more words to describe a novel concept. For example 'social media' is a new concept to PWN 3.1.

[7] See http://wordnetweb.princeton.edu/perl/webwn?s=post

Neologism type	Twitter	Reddit
Portmanteau	16 (14.8%)	4 (3.1%)
Novel Sense	18 (16.7%)	25 (19.2%)
Affixation	17 (15.7%)	16 (12.3%)
Phonetic	8 (7.4%)	9 (6.9%)
Loanwords	9 (8.3%)	7 (5.4%)
Compounds	15 (13.9%)	25 (19.2%)
Multiword Expression	12 (11.1%)	29 (22.3%)
Abbreviation	7 (6.5%)	5 (3.9%)
Other	6 (5.5%)	10 (7.7%)

Table 2. Breakdown of neologism construction methods for colloquial terms not in Princeton WordNet

Compounding Similar to above it is often common to create new words by combining two existing words into a new word, for example the combination of a 'hash' and 'tag' to make 'hashtag'. This is distinct from the previous category as it creates a single new word.

Affixation Many words are derived by adding a suffix or prefix to the word, in particular adding a prefix such as 're-', e.g., 'repost' or the affix '-ie' such as in 'selfie'.

Portmanteau The blending of two words to create a novel term, resulting in a word that contains phonetic characteristics of both words, such as 'cosplayer' (from 'costume' and 'player') or 'bromance' (from 'brother' and 'romance').

Loanwords Many novel words are loaned from other languages, examples include 'oppa' (from Korean) and 'waifu' (from Japanese).

Shortening Some novel words are created by shortening existing words, for example 'notif' (from 'notification') or 'sesh' (from 'session').

Phonetic spelling This is when neologisms are created by intentionally misspelling a word frequently for effect, for example 'smol' (from 'small') or 'bruv' (from 'brother'). It is also particularly common to see this in words that are associated with African-American Vernacular English, e.g., 'shawty' (from 'short').

Unknown For some words the derivation was not clear or could not be conclusively established, an example of this is 'twerk', whose etymology is unclear[8].

We classified the words into each of the categories and the results are presented in Table 2.

4 Publishing the Resource

We have made the data available under an open license, namely the Creative Commons Attribution (CC-BY 4.0) License in order to ensure that it can be

[8] http://blog.oxforddictionaries.com/2013/08/what-is-the-origin-of-twerk/

reused as widely as possibly. In addition, we have integrated our resource with two best practices in the area of WordNet data, namely with the Linguistic Linked Open Data Cloud and the Collaborative Interlingual Index from our website.[9]

4.1 Publishing the Resource as Linked Data

The Linguistic Linked Open Data cloud [4] has been proposed as a method for linking data between different resources and across modalities and these technologies promise to improve the interoperability and reusability of language resources on the Web. The OntoLex-Lemon model [5, 14] has been proposed as a model for the representation of lexical data on the Semantic Web and while its initial goal was to expand ontologies with lexical information it has recently been used for all kinds of lexical resources. We published the data using the Yuzu [15] system for linked data publishing and we link to the Polylingual WordNet [1], which has further links to the Interlingual Index. Using the Yuzu interface allows the data to be made available in RDF formats including Turtle, RDF/XML and N-Triples as well as JSON-LD [18].

4.2 Integrating the Resource with Collaborative Interlingual Index

The Collaborative Interlingual Index [2, 19] has been proposed as a method to enable cross-lingual development of wordnets. One of the major goals of this project has been defining a procedure by which new synsets can be defined and this goal overlaps with the objective of the Colloquial WordNet. Moreover, it is the case that non-English WordNets have not just introduced new concepts for words that are not directly lexicalisable in English, but have also introduced new synsets for novel concepts, often even when the term is a loanword from English. A notable example of this is the Polish plWordNet [13], which is significantly larger than any existing resource.

In order to facilitate the integration of Colloquial WordNet with the Collaborative Interlingual Index, we have made the full version of the resource available in the Global WordNet Association's recommended formats[10] and made it available under an open and permissive license. Furthermore, the Colloquial WordNet is participating in a pilot program to introduce the first set of new terms in the interlingual index and one term from the Colloquial WordNet, the verb 'to tweet', is a special test case as we believe this meaning is found in all major world languages.

5 Related Work

A previous project [7], called SlangNet, has already attempted to create a wordnet of slang for English, however this project has not released any version of the

[9] http://colloqwn.linguistic-lod.org/

[10] http://globalwordnet.github.io/schemas/

resource yet and appears to be inactive[11]. A certain number of our terms are also included in large-scale resources such as BabelNet [16] and we find that some of the terms added by our resource are already defined in BabelNet, however this is primarily only terms that are derived from Wiktionary and represents, 72.5% of our entries, which still means many terms would not be found in such resources. Similarly, the CROWN project [12] extended WordNet by means of automatically adding terms from Wiktionary.

The issue of detecting Neologisms has received some attention but approaches still have significant weaknesses. Neologism are of interest in traditional lexicography and major publishers work to detect neologisms [17] however these still rely significantly on manual work. Semi-automated detection has been attempted such as by extracting relevant features and classifying them using an SVM [8] or by relying on language-specific features [3]. We plan to use the training data we have collected in the first development round to improve the accuracy of the neologism collection procedure, using such supervised machine learning.

6 Conclusion

We have presented a method for development of a new extension to Princeton WordNet that covers the kind of language used in Twitter, Reddit and similar social media. Our resource relies on few annotators and as such we hope to encourage crowd validation by making the tool available online. Our extraction method relies on a mixture of corpus statistics and the usage of a crowd-sourced dictionary, Urban Dictionary, which we found to be of too poor quality to be used directly. While our method finds neologisms with high precision, we are not yet sure on the recall and as we cast a wider net this will become more critical. We analyzed the neologisms introduced and found that these terms are introduced not only by conventional methods such as affixation and sense extension, but also saw that a large number of words are entering as loanwords and in particular that portmanteaus are becoming much more common in colloquial English.

Acknowledgements

This work was supported in part by the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight) and NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1 TR000064 / UL1 TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH/NCATS.

References

1. Arcan, M., McCrae, J.P., Buitelaar, P.: Expanding wordnets to new languages with multilingual sense disambiguation. In: Proceedings of The 26th International Conference on Computational Linguistics (2016)

[11] We have aimed to combine this resource with our data, but discussions with the authors on licensing have been inconclusive.

2. Bond, F., Vossen, P., McCrae, J.P., Fellbaum, C.: CILI: the Collaborative Interlingual Index. In: Proceedings of the Global WordNet Conference 2016 (2016)
3. Breen, J.: Identification of neologisms in Japanese by corpus analysis. E-lexicography in the 21st Century: New Challenges, New Applications: Proceedings of ELex 2009, Louvain-la Neuve pp. 13–21 (2010)
4. Chiarcos, C., McCrae, J., Cimiano, P., Fellbaum, C.: Towards open data for linguistics: Linguistic linked data. In: New Trends of Research in Ontologies and Lexical Resources, pp. 7–25. Springer (2013)
5. Cimiano, P., McCrae, J.P., Buitelaar, P.: Lexicon model for ontologies: Community report. Final community group report, World Wide Web Consortium (2016)
6. Morgado da Costa, L., Bond, F.: Wow! what a useful extension! introducing non-referential concepts to wordnet. In: Proceedings of the 10th edition of the International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia. (2016)
7. Dhuliawala, S., Kanojia, D., Bhattacharyya, P.: SlangNet: A WordNet like Resource for English Slang. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation. pp. 4329–4332 (2016)
8. Falk, I., Bernhard, D., Gérard, C.: From non word to new word: Automatically identifying neologisms in French newspapers. In: LREC-The 9th edition of the Language Resources and Evaluation Conference (2014)
9. Fellbaum, C.: WordNet. Blackwell Publishing Ltd. (1998)
10. Grant, H.: Tumblinguistics: innovation and variation in new forms of written CMC. Master's thesis, University of Glasgow (2015)
11. Hicks, A., Rutherford, M., Fellbaum, C., Bian, J.: An analysis of wordnet's coverage of gender identity using twitter and the national transgender discrimination survey. In: Global WordNet Conference 2016 (2016)
12. Jurgens, D., Pilehvar, M.T.: Reserating the awesometastic: An automatic extension of the wordnet taxonomy for novel terms. In: HLT-NAACL. pp. 1459–1465 (2015)
13. Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S., Kedzia, P.: plWordNet 3.0–a Comprehensive Lexical-Semantic Resource. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. p. 22592268 (2016)
14. McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., et al.: Interchanging lexical resources on the semantic web. Language Resources and Evaluation 46(4), 701–719 (2012)
15. McCrae, J.P.: Yuzu: Publishing any data as linked data. In: ISWC 2016 Posters & Demonstrations Track (2016)
16. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)
17. O'Donovan, R., O'Neill, M.: A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. In: Proceedings of the 13th Euralex International Congress. pp. 571–579 (2008)
18. Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., Lindstrm, N.: Json-ld 1.1: A json-based serialization for linked data. Community group report, World Wide Web Consortium (2017)
19. Vossen, P., Bond, F., McCrae, J.P.: Toward a truly multilingual Global Wordnet Grid. In: Proceedings of the Global WordNet Conference 2016 (2016)