Towards a Crowd-Sourced WordNet for Colloquial English

John P. McCrae, Ian D. Wood
Insight Centre for Data Analytics
National University of Ireland Galway
Galway, Ireland
john@mccr.ae,
ian.wood@insight-centre.org

Amanda Hicks
Health Outcomes & Policy
University of Florida
Gainesville, FL USA
aehicks@ufl.edu

Abstract

Princeton WordNet is one of the most widely-used resources for natural language processing, but is updated only infrequently and cannot keep up with the fast-changing usage of the English language on social media platforms such as Twitter. The Colloquial WordNet aims to provide an open platform whereby anyone can contribute, while still following the structure of WordNet. Many crowd-sourced lexical resources often have significant quality issues, and as such care must be taken in the design of the interface to ensure quality. In this paper, we present the development of a platform that can be opened on the Web to any lexicographer who wishes to contribute to this resource and the lexicographic methodology applied by this interface.

1 Introduction

The Colloquial WordNet[1], first introduced in (McCrae et al., 2017), is an extension to Princeton WordNet (Fellbaum, 2010; Miller, 1995) that focuses on the use of neologisms and vulgar terminology[2]. The first version of this resource was created primarily by one lexicographer and as such scaling this resource to be able to cover more of the neologisms in English is a significant issue. In this paper, we detail the improvements we have made to the tools that lie behind this resource to enable a more open process for the creation of the resource. We started by detailing the guidelines and methodology for creating the resource and writing new documentation to support lexicographers in their work in annotating the data. We also added the possibility to add a confidence so that non-expert lexicographers would be able to provide annotations with some uncertainty. We then improved the interface in order to make it more intuitive for users with little knowledge of the project to use. In particular, we removed a lot of the 'implicit assumptions' of the interface that said that if certain options were chosen then other options could not be chosen. Furthermore, we integrated the guidelines in the editor so that lexicographers could easily look up the guidelines at any point where there is uncertainty. Finally, we introduced the idea of queues, where an annotator could add a number of terms, which have been automatically identified as potentially interesting, and these items can be held in the queue for a period of time, before being freed up. This methodology allows multiple lexicographers to collaborate without duplication of effort as each lexicographer's queue can be kept separate. The candidates that are in the queue are derived from Twitter and we detail the approach that we have taken to preprocessing the corpus and extracting the candidate terms from the result. Finally, we consider the issue of attracting new lexicographers for the resource and detail our plans to use student annotators and the creation of subtasks that may be of particular interest to individual lexicographers. These suggest a wider application of the methodology to more than just creating dictionaries for English neologisms. This project report represents the summary of recent work to create a resource that is more open and will be created by more than one lexicographer.

2 Colloquial WordNet Annotation Methodology

The methodology for creating Colloquial WordNet entries is based on annotating interesting words or short phrases from a corpus of tweets. The lexicographer will be presented with a lemma

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

[2] We are aware of a similar resource called SlangNet (Dhuliawala et al., 2016) but this does not seem to publicly available

and a number of example tweets and is expected to use these in order to write the entry. This is done in three steps: firstly, the lexicographer should check the lemma and examples and make sure he or she is familiar with the term or perform appropriate research in order to find the definition of the term. Then the lexicographer should sort the entry into its status (see Section 2.2), which will influence the method by- which it is further annotated. Then, the main body of the entry is created, in most cases in terms of the senses that define the meaning and any links to other senses.

2.1 Confidence

The first step in the creation of an entry in Colloquial WordNet is the selection of the lexicographer's confidence in the term. We decided to base these categories around the lexicographer's familiarity with the term, and the text guidelines are given as follows:

Very Strong : This is a term I use regularly and know exactly what it means (or the term is clearly an error, incomplete fragment of language or the name of a person, organization, etc.)

Strong : I am clear about the meaning of this term and have heard it used frequently

Medium : I have done a little research and am pretty sure I have a found a good definition

Weak : I have guessed from the term and the contexts and think I know what it means

Skip : I don't have a clue about this term and don't want to annotate it

Terms annotated with "skip" are returned to the queue for another lexicographer to handle. All other terms are included and the confidence can be used for other, more experienced lexicographers to check entries which may be weak.

2.2 Entry Status

The status indicates what kind of term this term is, note that "General", "Novel" and "Vulgar" are used for true terms, and "Abbreviation", "Misspelling", "Name", "Not Lexical" and "Error" for terms that will only be included in the ancillary data for Colloquial WordNet.

General : This is a term that should be included in a general-purpose dictionary such as Princeton WordNet. It should be widely and frequently used by native English speakers. Example: "lockpick"

Novel : This term is novel and may not persist in the language. This term should be used for slang, dialectal forms (used only in a particular dialect or social group) and other non-standard usage of English. This should also be used for interjections such as "wow!" or "gosh!" (in this case, the part of speech should be other). Examples: "twerk", "dab", "belieber"

Vulgar : This term is vulgar or obscene and would not be suitable for a general purpose dictionary. Examples: "mindfuck", "paypig".

Abbreviation : This term is an abbreviation; Examples: "IDK", "IMHO"

Misspelling : This term is misspelled; Examples: "agnst", "newjob"

Inflected Form : This term is an inflected form, a simple grammatical variation of a word (e.g.: "running" from the word "run"). Examples: "cats", "the cat"

Name : This term is a name (proper noun) and is not suitable for inclusion in the WordNet. Examples: "Google", "Justin Bieber"

Not Lexical : This is not a proper term. It may be a fragment of text that doesn't make sense as an independent phrase, e.g., "I know a", or it may be a multiword phrase, where the meaning is clearly composed from the constituent words, e.g, "tasty ham", "cheese sandwich".

Error : This is used if the "term" does not seem to be English, e.g., " "

2.3 Entry Details

The entry details are the main work for the lexicographer. For entries, whose status is "General", "Novel" or "Vulgar", the lexicographer will enter the senses as either novel senses with definitions and relations or as synonyms of existing WordNet entries, for which an auto-suggest feature is used to help the lexicographer. This allows the lexicographer to type the lemma of the synonym and then

they are shown the part-of-speech, definition and an Interlingual Index (ILI) ID (Bond et al., 2016; Vossen et al., 2016). In the case the lexicographer chose either "Abbreviation", "Misspelling" or "Inflected form" the lexicographer simply fills in the lemma that should be used here, i.e., the unabbreviated, noninflected, correctly spelled word. For misspellings and inflected forms this lemma is then queried against existing PWN and Colloquial WordNet entries and if it is not found then it is re-added with the correct lemma to the user's queue. We require that each new word has at least one link, this is generally to an existing synset in Princeton WordNet, through the Interlingual Index (Vossen et al., 2016; Bond et al., 2016), however it may just be to another existing Colloquial WordNet entry, e.g., "retweet" and "subtweet" to "tweet".

3 Building an interface for crowd-sourcing

In order to support lexicographers in creating their interface, we have designed an attractive user interface (see Figure 1), that can be used to create new entries in the Colloquial WordNet. The interface is created using Scalatra[3], is backed by an SQLite Database[4] and uses Bootstrap[5] and Angluar[6] for the user interface. These technology choices were made in order to create an interface with reduced effort.

3.1 Queues

Queues are the main interface that a lexicographer uses to select the terms that they wish to annotate. The lexicographer can choose to add elements to their queue, and these are taken from the most important terms that have not yet been annotated. Once they are entered into the queue they are locked and can only be annotated by this lexicographer for the next 7 days. Lexicographers may remove or extend terms from their queue, and in editing mode, once a lexicographer submits an entry the website automatically redirects them to editing the next entry in their queue or back to the queue page if there are no elements left in their queue.

[3] http://scalatra.org
[4] https://www.sqlite.org/
[5] http://getbootstrap.com/
[6] https://angularjs.org/

3.2 Tweet Collection and Preprocessing

In order to get a sample of current social media language usage, we have been collecting tweets from the "sample" endpoint of the public Twitter streaming API. This provides a continuous stream consisting of a 1% sample of all published tweets. Collection has been ongoing since August 2016, resulting in 435 million English language tweets as of August 2017.

In an attempt to reduce the impact of unintelligible tweets, robots and spam, we apply the following simple rules:

Small Words : We remove tweets if they contain lots of short words. A short word is defined as a word with 1 or 2 characters and we remove the tweet if more than 30% of words are short.

New Lines : We remove tweets with more than two newlines, as these are likely to be advertising or spam.

Non-dictionary words : We check all words against a dictionary of known English words and reject tweets where more than 30% are not in the dictionary. This removes tweets not in English.

Tags : We count the number of words starting with a '#' or '@' and remove it if more than 30% of words start with such a tag.

Tweets matching these rules were mostly either not expressions of natural language or identified as automatically generated tweets. Applying these heuristics substantially reduced the number of tweets, resulting in a collection of 34,776,298 tweet texts as a sample of contemporary social media language usage.

An important feature of this collection is that it spans a whole year. This reduces the effects of high word frequencies associated with specific content associated with large social media coverage (coverage of events such as sports matches, elections, annual television events etc. . . or tweets that "go viral"). The ongoing and longitudinal nature of the data also permits analysis of *changes* in language usage over time, a topic we intend to investigate in future work.

3.3 Selecting Candidates

Once we have identified the tweets, we attempt to find the words that are most relevant to be anno-

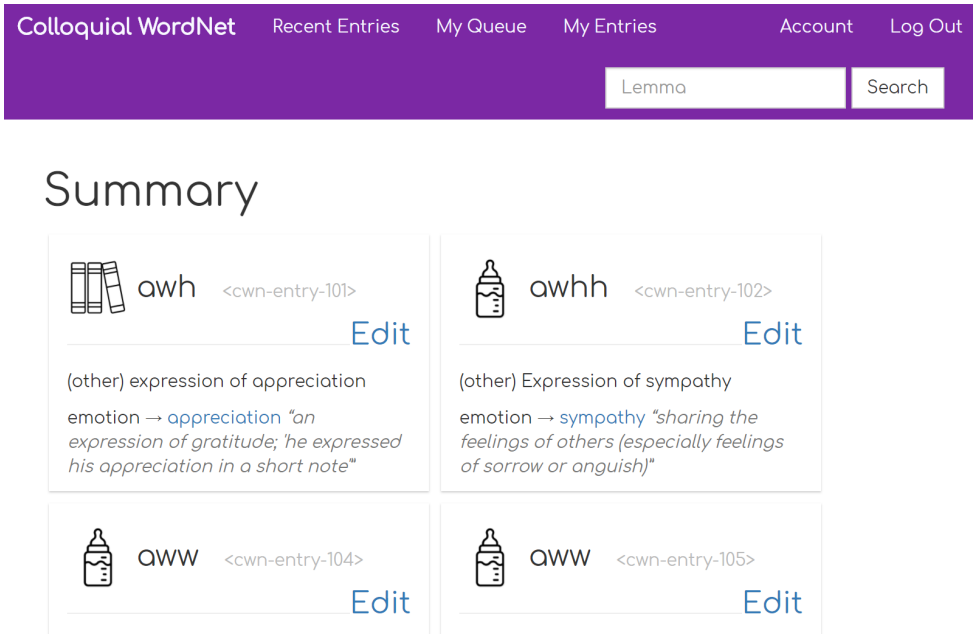

Figure 1: The Colloquial WordNet Editor Interface

tated. For this, our primary approach is to use the frequency relative to a background corpus, in particular from a Web Corpus of term frequencies [7]. Our approach chooses the terms that have a high frequency relative to the baseline corpus and in addition we choose terms that mostly occur in all lowercase to remove many of the proper nouns and other terms that are present in tweets. For each of the selected terms we also choose 10 example tweets to help the annotator, these are chosen based on a variation of the GDEX algorithm (Kilgarriff et al., 2008), where in particular we rank tweets based on:

Length If the tweet is between 10 and 25 words.

Blacklisted Words Whether the tweet contains any blacklisted words, such as 'this', 'that' or 'http'

Punctuation Whether the tweet starts with a capital letter and ends with a full stop, question mark or exclamation mark.

Frequent Words How many of the words in the tweet are in the top 17,000 words.

These tests give each tweet a score out of 4, with 'frequent words' used as a tiebreaker. We greedily choose the top 10 example tweets, in addition requiring that no tweet overlaps by more than 5 words with a previously selected tweet.

[7] http://norvig.com/ngrams/

4 Supporting Lexicographers

To facilitate the development of Colloquial WordNet future work involves using linguistics students as annotators and creating subtasks focused on a particular domain of interest so that lexicographers who are proficient with the use of terms that are specific to particular subdomains and communities.

4.1 Gender minority and Pro-Ana Subtasks

We have developed two specialized Twitter corpora in previous projects (Hicks et al., 2015; Wood, 2015), that can also be used to find domain specific terms for addition to Colloquial WordNet and to attract annotators who are interested in and drawn to a specific topic. One corpus, first reported in (Hicks et al., 2015), represents tweets over a period of 49-day period from January 17, 2015 to March 6, 2015 inclusive that contain terms related to gender identity, particularly terms that indicate a transgender or other gender minority identity, (e.g., "transboi", "FTM", and "non-binary"). A pilot interface has been created around this corpus using the method described in the previous section to suggest candidate terms for inclusion in the Colloquial WordNet.

The second Twitter corpus, originally report in (Wood, 2015), represents tweets over a period of nearly three years (December 2012-October 2015) that contain hashtags that may indicate

membership of the "pro-anorexia" and eating disorder community (e.g., "#proana", "#edproblems", and "#thinspiration").

While these domain specific subtasks contain community specific neologisms, they also contain general terms that may not already be included in WordNet (e.g. "trans woman" and "queerness"). Many of the candidate terms derived from the gender minority corpus are not specific to gender identity (e.g. "tummy tuck", "woc" as an abbreviation for woman of color and "tranny" as a synonym of "transmission"). Furthermore, a coverage analysis of WordNet's gender identity terms showed that adding a small number of wordsenses to WordNet can result in significantly greater coverage of gender identity terms in WordNet due to the prevalence of compositional multi-word expressions used to describe gender identity. (Hicks et al., 2016). We anticipate that these subtasks will also increase coverage of non-domain specific terms while retaining the interest and participation of annotators who are drawn to the topic.

5 Conclusion

In this paper we present the progress in the development of Colloquial WordNet editor and its tools. While there exist many other tools for editing WordNets, e.g., DebVisDic (Horák et al., 2006), SlowTool (Fišer and Novak, 2011), plWNApp (Derwojedowa et al., 2008) or Wordnetloom (Piasecki et al., 2013), none of these tools meet our goal of being an open Web-based development platform that can be used by any user. The goal of Colloquial WordNet is to be more open, and as such we do not necessarily expect the same level of expertise from our lexicographers or quality in the resulting resource. Instead, we understand Colloquial WordNet to provide a good WordNet-level coverage of English as it used in social media, which will be helpful to handling noisy user-generated text, a problem that has caused significant issues for natural language processing recently (Baldwin et al., 2015). Currently the resource consists of the same 428 entries previously detailed (McCrae et al., 2017), however we now expect to work on expanding the resource. Furthermore, we believe that the exercise of developing the Colloquial WordNet can identify key words that we hope will contribute to the next version of Princeton WordNet and should assist the lexicographers by providing entries that can be further extended into PWN entries.

Acknowledgments

This work was supported in part by the Science Foundation Ireland under Grant Numbers SFI/12/RC/2289 (Insight) and 16/IFB/4336 and also in part by the NIH/NCATS Clinical and Translational Science Award to the University of Florida UL1 TR000064. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the NCTE.

References

[Baldwin et al.2015] Timothy Baldwin, Young-Bum Kim, Marie Catherine De Marneffe, Alan Ritter, Bo Han, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *ACL-IJCNLP*, 126:2015.

[Bond et al.2016] Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*.

[Derwojedowa et al.2008] Magdalena Derwojedowa, Maciej Piasecki, Stanisław Szpakowicz, Magdalena Zawisławska, and Bartosz Broda. 2008. Words, concepts and relations in the construction of Polish WordNet. In *Proceedings of the Global WordNet Conference, Seged, Hungary*, pages 162–177.

[Dhuliawala et al.2016] Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Slangnet: A wordnet like resource for english slang. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, pages 4329–4332.

[Fellbaum2010] Christiane Fellbaum. 2010. WordNet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

[Fišer and Novak2011] Darja Fišer and Jernej Novak. 2011. Visualizing sloWNet. *Proceedings of the Electronic Lexicography in the 21st Century (eLex 2011)*, pages 76–82.

[Hicks et al.2015] Amanda Hicks, R. Hogan, William, Michael Rutherford, Bradley Malin, Mengjun Xie, Christiane Fellbaum, Zhijun Yin, Daniel Fabbri, Josh Hanna, and Jiang Bian. 2015. Mining Twitter as a first step toward assessing the adequacy of gender identification terms on intake forms. In *Proceedings of the AMIA 2015 Annual Symposium*. American Medical Informatics Association.

[Hicks et al.2016] Amanda Hicks, Michael Rutherford, Christiane Fellbaum, and Jiang Bian. 2016. An

analysis of WordNets coverage of gender identity using Twitter and the national transgender discrimination survey. In *Proc of the Eighth Global WordNet Conference (GWC)*, volume 2016, pages 122–129.

[Horák et al.2006] Aleš Horák, Karel Pala, Adam Rambousek, and Martin Povolnỳ. 2006. DebVisDic–first version of new client-server wordnet browsing and editing tool. In *Proceedings of the Third International WordNet Conference–GWC 2006*, pages 325–328.

[Kilgarriff et al.2008] Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of Euralex*.

[McCrae et al.2017] John P. McCrae, 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)*.

[Miller1995] George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

[Piasecki et al.2013] Maciej Piasecki, Micha Marciczuk, Radosaw Ramocki, and Marek Maziarz. 2013. Wordnetloom: a Wordnet development system integrating form-based and graph-based perspectives. *Int. J. Data Mining, Modelling and Management*, (5).

[Vossen et al.2016] Piek Vossen, Francis Bond, and John P. McCrae. 2016. Toward a truly multilingual Global Wordnet Grid. In *Proceedings of the Global WordNet Conference 2016*.

[Wood2015] Ian Wood. 2015. A case study of collecting dynamic social data: The pro-ana Twitter community. *Australian Journal of Intelligent Information Processing Systems*, 14(3).