

SlangNet: A WordNet like Resource for English Slang

Shehzaad Dhuliawala, Diptesh Kanojia, Pushpak Bhattacharyya

Indian Institute of Technology Bombay
Mumbai, India
{shehzaadzd, diptesh, pb}@cse.iitb.ac.in

Abstract

We present a WordNet like structured resource for slang words and neologisms on the internet. The dynamism of language is often an indication that current language technology tools trained on today's data, may not be able to process the language in the future. Our resource could be (1) used to augment the WordNet, (2) used in several Natural Language Processing (NLP) applications which make use of noisy data on the internet like Information Retrieval and Web Mining. Such a resource can also be used to distinguish slang word senses from conventional word senses. To stimulate similar innovations widely in the NLP community, we test the efficacy of our resource for detecting slang using standard bag of words Word Sense Disambiguation (WSD) algorithms (Lesk and Extended Lesk) for English data on the internet.

Keywords: Slang, WordNet, Informal Text, Lexicon

1. Introduction

The internet is an ardent platform for users to interact. It has given rise to several public forums and bulletins where users communicate with each other. The fact that the internet is global, allows for this userbase to not be confined to a particular region or location. Users from different countries, of different races and languages can speak with each other. This vast diversity has allowed language to morph on the internet. One often notices words, phrases and colloquialisms which pop into existence and are used in huge volumes over the internet. The following examples better illustrate the need to annotate novel senses for the following words:

Sick (Adj): Used in a positive sense to denote something that is nice or awesome. *Eg.* The band's new album is **sick**.

Bae (Noun): Used to describe something or someone you like or find attractive. Also used to describe your romantic partner. *Eg.* Don't you hate it when your **bae** flirts with another guy.

Text (Verb): Sending a short message over a mobile device. *Eg.* Can you **text** me when you reach home?

This motivates us to create a resource that could contribute to various NLP applications which work on noisy data collected from the internet. We utilize a popular online resource, called Urban Dictionary as a reference, to create SlangNet.

Urban Dictionary is an online, crowd sourced slang dictionary. It allows users to input definitions and examples for common slang words on the internet. The website has a user upvote and downvote system which decides the popularity of the definition. The site has a user base of 18 million unique readers and hosts about seven million words and phrases along with their definitions¹. Due to no internal or external validation, Urban Dictionary fails to control

the quality of the definitions. Swerdfeger (Online)² observe that the resource suffers from redundancy, self-references, opinions of users, spelling errors, grammatical errors, contrasting definitions *etc.* The shortcomings of using Urban Dictionary in its very raw form make it unusable as a lexical resource.

Princeton WordNet or the English WordNet (Fellbaum, 1998) is an online lexical resource which can be used for various NLP applications such as WSD, Machine Translation, Information retrieval, etc. Based on English WordNet, several other WordNets like the EuroWordNet(Vossen, 1998), IndoWordNet(Bhattacharyya, 2010) and MultiWordNet(Pianta et al., 2002) were created. We create a WordNet like structure which can be utilized for the aforementioned NLP applications. Other such works which are built upon a WordNet like structure and produce augmented resources are BabelNet(Navigli et al., 2010) and FrameNet(Baker et al., 1998). VerbNet(Schuler, 2005) is another such verb lexicon currently available for English. It is a hierarchical domain-independent, broad-coverage verb lexicon with mappings to the English WordNet. ConceptNet(Liu and Singh, 2004) is a semantic network, built from nodes (or "terms"), representing words or short phrases of natural language, that labels the relationships between them.

Our method helps refine the data from Urban Dictionary; using several manual changes and mappings, we normalize this to a structured WordNet like resource. We validate the usability of our resource by implementing conventional unsupervised Word Sense Disambiguation (WSD) methodologies such as Lesk(Lesk, 1986) and Extended Lesk(Banerjee and Pedersen, 2003) algorithms.

Our resource aims to facilitate NLP tools with a properly constructed set of definitions to deal with slang on the internet. Several applications try to harness the data on the internet as corpora, but most of them treat slang words as noise. Many slang words do contain information which could help

¹https://en.wikipedia.org/wiki/Urban_Dictionary

²<http://www.cs.ubc.ca/~carenini/TEACHING/CPSC503-14/FINAL-REPORTS-07/BradSwerdfeger-FinalPaper-1.pdf>

improve results for various NLP applications. Our resource would enable other researchers to harness this information.

2. Creating SlangNet

The following section describes the pipeline for generating SlangNet and refining its data.

2.1. Types of slang words

Slang words fall into two categories:

Newly created words It includes words which don't presently exist in the English dictionary. These words are entirely new. Example: *swag* def.: style, *turnt* def.: drunk *etc.*

Newly created senses This category includes words which do exist, but are commonly used with a different sense. Example: Cool (The best way to say something is neat, awesome, or swell) rather than the traditional sense: the quality of being at a refreshingly low temperature.

2.2. Identifying slang words

We crawl comments from Reddit³ and create a corpus containing about 5.4 million sentences. We then iterate over the words present in these comments to bootstrap our resource. Words which appear multiple times and are not present in the WordNet pose a high probability of being a neologism or a slang word. We then query the Urban Dictionary API⁴, and search for possible senses of the word. The top definition is picked up and its gloss and example(s) are chosen. Urban Dictionary also provides a set of similar word tags, which are also selected to better identify the correct synset for the word.

For slang words which fall into the second category, a more involved approach is followed: For content words, in the text, Urban Dictionary is queried and the tags of the words are retrieved. We find that these tags contain words similar to the all the senses (conventional and slang). For example, for the word *sick*, the tag words are: ("awesome", "cool", "ill", "sweet", "sickness", "amazing", "gross"). One can observe that this set contains words relating to the traditional definition of the word *sick*(ill), along with the more novel definition(awesome). We use the following method to identify new senses: A basic bag-of-words based approach is used to score each of the tags. The score is calculated on the basis of an overlap between the tag and the gloss of each sense of the word. We try to match each tag to the conventional senses in the aforementioned manner. If a particular tag fails to match any of the WordNet senses, we try to match it to the gloss from Urban Dictionary. If a match occurs, then we choose the tag word to represent the slang sense of the word.

2.3. Refining and structuring

Very few of the definitions provided by Urban Dictionary are close to conventional. The primary reason for sub-standard quality of definitions is the lack of moderation. We attend to refine this information by manual intervention.

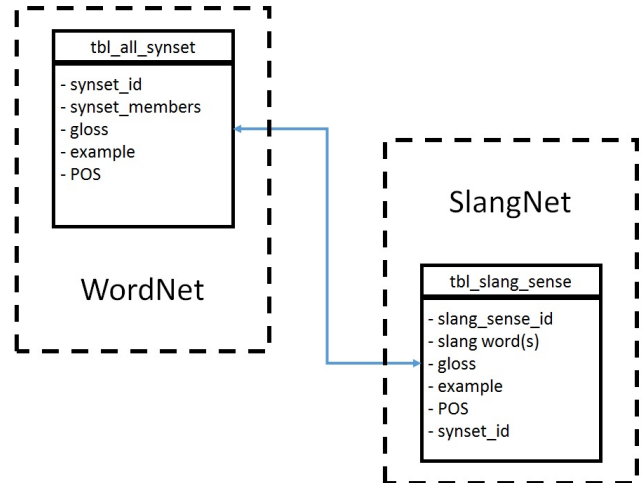


Figure 1: UML Data Model Diagram for SlangNet

Word	simple Lesk	extended Lesk
Cool	0.15	0.17
Insane	0.38	0.54
Own	0.45	0.51

Table 1: Results for detecting slang senses

A set of human annotators assign the correct definition where Urban Dictionary fails. The annotator chooses the correct synset to be linked to the slang word. If no possible match is available, a new synset is created for the same. To ease the annotation process, we created a tool (described in section 4.), using which, the annotator is given a list of suggested WordNet senses for the slang word along with the definition obtained from Urban Dictionary. When a definition needs to be assigned, the definition from Urban Dictionary is only used as a reference and a more formal definition is assigned.

We perform an initial automated POS tagging on the corpus. However, owing to the noise in the data, the accuracy of this method was experimentally found to be lower than the general accuracy. To mitigate this, the annotators also assign the specific POS tag for the slang word after considering its usage in text. Several slang words (especially acronyms such as 'LOL' (Laughing Out Loud) or 'BTW' (By The Way) cannot be assigned a POS tag. An 'ABR' tag was assigned to these words to indicate that they are abbreviations.

2.4. Dynamism

The purpose of this resource is to keep up with the continuously updating language of the internet. For this reason, a static resource for slang would fail its very purpose once today's slang goes out of use. We aim to create a web-crawler which runs continuously on various user forums and bulletins. The crawler would monitor the usage of words over these websites. A new word which appears poses the probability of being a slang word. Further we also aim to use methods such as described in Cook et al. () and Lau et al. (2012) to monitor if a new sense seems to be emergent.

³<http://www.reddit.com>

⁴<http://api.urbandictionary.com/>

Word	Gloss / Definition	Example	POS
Text	A short message sent using a mobile device usually through a protocol such as a short messaging service (SMS)	I received a very rude text on my mobile	Noun
Poster	A person who uploads a message or comment online on a social media website	The poster was banned for a vulgar comment	Noun
Follower	someone who follows your activity on a social media website or application	He has millions of followers on Twitter	Noun

Table 2: New senses from Cook et. al. dataset

Word	simple Lesk
Follower	0.31
Poster	0.19
Text	0.42

Table 3: Results for detecting slang senses

3. Validation

We validate our claims of a structured, usable resource by performing experiments using WSD engines. We use the Lesk and Extended Lesk algorithms, implemented through PyWSD(Tan, 2014) to perform WSD on 6 chosen slang words (from two datasets). For both these experiments we use our resource as an addendum to the current WordNet.

3.1. Validation Dataset

3.1.1. Reddit Corpus

We choose 100 sentences randomly for each of the following 3 words: Cool, Own, Insane. We perform manual validation over those 100 sentences by manually tagging the correct sense from WordNet + SlangNet. While tagging, we observe 97 % occurrences for the word ‘cool’ belonging to the sense added in SlangNet. Similarly, we also observe 56 % occurrences for the word ‘own’, and 88 % for the word ‘insane’ belonging to their respective SlangNet senses⁵. Two human annotators tagged the occurrences for the word ‘insane’ by randomly picking 50 contiguous sentences from the corpora. Each occurrence was marked with its sense tag. The two human annotators have had 15+ years of academic instruction in English. We observe their inter-annotator agreement via kappa score to be 0.84.

3.1.2. Novel sense Dataset

Here we use the dataset provided by Cook et al. (). This is a pre-annotated dataset, consisting of new senses from the computing domain. We choose 3 words relating to and regularly used in social media. We use the Lesk algorithm along with our resource to judge its efficacy on this dataset. We do not experiment with Adapted Lesk as we are still in the process of adding semantic relations for our resource. The words and their assigned definitions are given in Table 3.1.2.

3.2. Analysis

The accuracy obtained via Lesk and extended Lesk algorithms are reported in Table 1, which are more or less the

⁵**Cool**: Awesome, **Insane**: Crazy, fantastic, **Own**: defeat someone (mostly in an online game)

same as reported in previous works (Banerjee and Pedersen, 2003). This shows that SlangNet when integrated with the WordNet allows WordNet based algorithms such as the Lesk to seamlessly run and disambiguate senses.

We believe that the lower accuracy for the word ‘cool’ is because it holds several fine grained senses, thus increasing sparsity.

4. SAnE: Slang Annotation and Evaluation

We develop a tool for slang word annotation, which takes as input the raw retrieved text from Urban Dictionary. The annotator is shown the definition from Urban Dictionary, an example of the word’s usage from Urban Dictionary and from the corpus. Based on the Urban Dictionary definition and tags, a set of conventional synsets are predicted. The annotator may choose to do one of the following:

- Choose an appropriate conventional synset (from the predicted list, or manually) for the slang word.
- If the meaning fits no conventional sense, a new SlangNet synset is created and the annotator fills in a definition, POS tag and example(s).

5. Conclusion and Future Work

We present a WordNet like resource which can augment the English WordNet while dealing with neologisms and slang words on the internet. We show that the general accuracy obtained by using Lesk and extended Lesk is greater than when Urban Dictionary is directly used (Swerdfeger, Online). Currently, our resource holds 3000 slang words. However, this figure is constantly being updated as described in section 2.4. above. Our resource aims to mitigate the effect of non-conventional language on the internet.

We notice that several of these slang words are sentiment bearing and annotating them with sentiment scores like the SentiWordNet (Esuli and Sebastiani, 2006) would help Sentiment Analysis for data on the web. We aim to do the same as a part of our future work. Also as a future work, we aim to span out to different languages. We plan to identify slang from various languages and link semantically similar slang words. For example, the word LOL (Laughing Out Loud) is synonymous to the French MDR (Morte De Rire). Such linkages could be used to aid real-time Machine Translation in both text-to-text and speech-to-text scenarios.

We aim to continuously update our resource with respect to adding new slang words to our repository. We also aim to make our resource, implementation and tools publicly available for the research community. (Speecon Consortium, 2014).

6. Bibliographical References

- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Banerjee, S. and Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI*, volume 3, pages 805–810.
- Bhattacharyya, P. (2010). Indowordnet. In Nicoletta Calzolari, et al., editors, *LREC*. European Language Resources Association.
- Cook, P., Lau, J. H., McCarthy, D., and Baldwin, T.). Novel word-sense identification.
- Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Cite-seer.
- Fellbaum, C. (1998). *WordNet*. Wiley Online Library.
- Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.
- Liu, H. and Singh, P. (2004). Conceptnet-a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.
- Navigli, R., Roma, S. U. D., and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *In Proc. of ACL-10*.
- Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.
- Schuler, K. K. (2005). Verbnet: A broad-coverage, comprehensive verb lexicon.
- Swerdfeger, B. A. (Online). Assessing the viability of the urban dictionary as a resource for slang.
- Tan, L. (2014). Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software]. <https://github.com/alvations/pywsd>.
- Piek Vossen, editor. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

7. Language Resource References

- Speecon Consortium. (2014). *Dutch Speecon Database*. Speecon Project, distributed via ELRA, Speecon resources, 1.0, ISLRN 613-489-674-355-0.