

Assessing the Viability of the Urban Dictionary as a Resource for Slang

©Bradley A. Swerdfeger. All rights reserved.

Bradley A. Swerdfeger
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada
bswerd@cs.ubc.ca

Abstract

The use of slang is ubiquitous, especially in internet communities. This paper evaluates the success of conventional dictionary and thesaurus-based semantic similarity assessments on The Urban Dictionary, an online, user-contributed dictionary for contemporary slang. Conventional methods are shown to perform poorly, and problematic aspects of the corpus are examined. Language use on the internet is found to be very unconventional, and techniques designed for conventional, well-formed language are likely to perform poorly. Future work is suggested in order to understand unconventional language use and the development of neologisms.

1 Introduction

The use of slang is ubiquitous in colloquial speech. Even the most refined of individuals will use slang whilst speaking with family, friends, and sometimes co-workers. Slang is also present within popular media such as newspapers and magazines and it is used excessively in movies and television. Due to the nature of human language, slang is causing an explosion of neologisms and synonyms; new words are simply being invented every day, especially within the contexts of conversational speech and the internet. However, slang and colloquialisms have been largely ignored by the linguistic community, as it is seen as a language used by uneducated and lower-class individuals (Millhauser, 1952). WordNet contains only the most well-established and undeniable uses of slang, but fails to capture more recent developments. In order to understand human language usage in most contexts, we must have a computational understanding of slang.

What separates slang from conventional language? Cooper (2005), revealed that the number of senses per word in a dictionary has a near-exponential distribution. This means that most words only have a single definition, and as the number of definitions for a word increases, the number of words with this number of definitions decreases exponentially. However, this is not the case with slang. With slang, the distribution is nearly flat, with a slight decrease as we approach a greater number of senses per word. Slang is, therefore, highly ambiguous in meaning. Note that slang was not the focus of this study; it

was just used as a comparison to conventional English and French.

This ambiguity is problematic for natural language understanding in applicable domains. For instance, WordNet has 5 senses for the word *wicked*, all of which have very negative connotations (synonyms include *sinful*, *terrible*, *prankish*, or *disgusting*). However, *wicked* is used as a positive evaluation or denotes emphasis when used in the context of slang. This complete reversal of polarity is problematic for natural language understanding, especially in the context of online evaluations.

Without an accurate and current corpus, dictionary, or thesaurus for slang, traditional word sense disambiguation techniques cannot be employed. Obtaining these entities is an especially difficult task due to the rapid evolution of slang. Although Webster does publish a dictionary for slang, it can take a very long time for a term to enter its pages. Also, its coverage is quite limited and does not accurately represent the explosion of neologisms due to slang. The most feasible way of collecting this information is to have the slang-users themselves provide the dictionary.

The Urban Dictionary is a web-community that attempts to do this. It is an online dictionary where users can submit definitions for words or even create new words with a provided definition. As of December 2005, it boasted over 300,000 definitions and was growing at a rate of 2,000 definitions per day. Users can rate a definition with either a ‘thumbs-up’ or ‘thumbs-down’, so that the more popular and accurate definitions are highlighted. One can see it as a dictionary for contemporary slang since it captures most of the neologisms being used in every day conversation.

Given that this resource is uncontrolled and somewhat organic in its generation, to determine if it can actually be used, we must first assess its viability for use in conventional computational tasks. One such task that is useful for word sense disambiguation is calculating semantic similarity using dictionary methods in order to assign a term to an existing synset. This type of task would be useful in order to augment resources such as WordNet. Given the conventional nature of this task, performance might give us a good idea of how well techniques developed for conventional language work for unconventional language use. This is the deeper and more important research question.

This paper first describes related work (Section 2), then describes my contributions to this task (Section 3). The corpus (Section 4) and gold standard (Section 5) are described and then a high-level description of the implementation used to

test the viability of the corpus is given (Section 6). Results are presented in Section 7, and an extensive error analysis is conducted to determine issues with the corpus (Section 8). Revisions are made to the gold standard, and the new results are presented (Sections 9 and 10, respectively). Lessons learned from completing this work are presented in Section 11, and then an evaluation of the success of the project (Section 12) and suggestions for future work are given (Section 13).

2 Related Work

To the author's knowledge, this is the first instance of work in linguistics that deals with The Urban Dictionary. The author is also not aware of any work in computational linguistics that attempts to deal with the use of slang or semantic similarity in unconventional domains. There is, however, a great deal of work on the subject of conventional semantic similarity using dictionaries or thesauri.

Kilgarriff and Rosenweig (2000) introduced the LESK algorithm for word sense disambiguation tasks, and it is often used as a baseline for the problem. This method computes the overlap of words between glosses; therefore it can be used for semantic similarity as well with the overlap revealing similarity.

Vasilescu (2004), extended the LESK algorithm to include a weighting of overlap by the inverse document frequency of the overlapping words. This method gives more weight to less frequent words, and therefore assesses similarity based on the words that give the most new information.

Lin (1998) introduces an information theoretic definition of similarity that is related to Resnik similarity (1995). This method assesses similarity as twice the information content of the lowest common subsumer divided by the information content of the sum of the words in question. This measure extends Resnik by making the observation that similarity assessments must measure more than the information in common between two words. They should also measure the differences between two words, which indicate how dissimilar they are. His similarity measure, then, describes semantic similarity as the ratio between the amount of information needed to state the commonality between two words and the information needed to describe what the two words are.

Finally, Curran's PhD thesis (2003) describes a method of measuring association with context based on the t-test statistic. This method assumes that the null hypothesis is that two words are independent, and it measures the difference between observed and expected means, normalized by the variance. Here, the expected mean is the null hypothesis. A high value of this measure shows that the association between words is higher than can be expected by chance. This association measure is used as a feature value in vectors that describe the context of words. The t-test statistic is computed between a word and the words found within a window around the target word. Similarity is then assessed by comparing the feature vectors to find which words appear in the similar contexts.

These methods will all be used for assessing semantic similarity within this paper in order to assess The Urban Dictionary as a resource that will work using conventional methods.

3 Contributions

This work provides a first look at using only online resources for computing the semantic similarity of slang terms. The online resources are assessed, and weaknesses are presented. In light of these assessments, directions for future work are presented in order for the computational linguistics community to develop a better understanding of slang for the use of computational techniques in more natural or colloquial domains.

4 Corpus

The corpus was collected by implementing a link-following web spider and deploying it on www.urbandictionary.com. The spider parsed the first page of definitions for each word in the dictionary, including the example of usage given and the ratings of the definitions. The definitions are ordered by votes; therefore the most popular definitions are situated on the first page.

After running the spider for two weeks, 627 138 definitions were collected, with the spider being terminated at *slo...* due to a major change to the layout of The Urban Dictionary that was made while it was being parsed. An example definition that represents the information collected can be seen in (1).

Term: *manther* (1)

Definition: *A male cougar. Single, usually divorced, and at a minimum 10 years older than a cougar.*

Example: *He did not care that the youth laughed at his ragtop corvette that even at his age he could not afford, for he was a manther.*

Likes: 350

Dislikes: 160

As you can see in (1), there is no part-of-speech or sense annotation given to the definitions. Therefore, multiple definitions for the same sense or different senses can occur on the first page, without any *a priori* way of disambiguating them; in fact, this is the norm.

Given this dearth of information, only the highest rated definition for a term was kept. Furthermore, only terms for which a match was found in the gold standard survived. Finally, only terms whose Likes to Dislikes ratio is over 2.0 and terms that have over 20 Likes were kept. This means that the highest rated definition had to be rated unambiguously positive in order to be used for implementation and testing.

This filtering reduced the number of definitions to 1074. No regard was given to the sense of the terms in the filtering since this information is unknown.

The reduction of 627 138 definitions to 1074 means that only 0.17% of the collected definitions were actually used. The remaining 626 064 definitions could not be used without any human intervention in the form of sense annotation or addition to the gold standard. Given such a large number, obviously, this task is not feasible.

5 Gold Standard

For the initial assessment of The Urban Dictionary as a resource for slang, I wanted to use only existing resources as a substitute for annotation. The reasoning behind this was three-fold. First, 627 138 definitions was quite daunting, and time constraints did not allow for me to annotate any respectable portion of this number. Secondly, conducting a user study to assess subjective semantic similarity is only feasible for a relatively small number of terms (Miller and Charles, 1991). With such a large number of available definitions, it would be very difficult to select 30 to 40 definitions where several are semantically very similar. Finally, given the widespread use of slang, there is a need to scaffold existing techniques by using existing resources in order to see benefits as quickly as possible.

Due to these reasons, a more controlled, but less extensive and popular online resource for slang was used. OnlineSlangDictionary.com is a resource where website visitors can submit definitions for slang terms, but they are edited and approved by the site moderator. The dictionary contains 1412 terms, annotated with one definition per sense, multiple examples per sense, parts of speech, and related categories of words (2). There are 1855 definitions in total. As you can see, this dictionary has much less coverage than The Urban Dictionary. It has also not been updated since August 2006, which eliminates it as a candidate for a current and accurate resource for slang. However, it can serve as a gold standard with a decent size as it is somewhat recent and contains some very popular uses of slang.

Term: *melon* (2)

POS: *Noun*

Definition 1: *A head.*

Example 1: *I hit my melon playing baseball.*

Definition 2: *moron, idiot.*

Example 2: *Jim is such a melon.*

Related: *{head},{unintelligent}*

To define the gold standard, only the *Related* field was used. This represents categories under which a term might belong. For instance, synonyms for drugs will have a *Related* field of *{alcohol, drugs, tobacco}*. A word can have more than one 'synset', as is shown in (2). Although these relationships are not exactly synsets, they were used for two reasons. First, there is some sense of semantic similarity within these categories. They will contain words that are related, although they may not be synonymous. This also means that there will likely be a synonym within the set; therefore if we can find the closest word, we can assign a category to a term with some certainty. Secondly, the categories already exist. No further annotation must be performed.

This process discovered 353 different synsets, with multiple synsets containing instances of the same word. Synsets contained an average of 5.2 words each. The largest synset was *{cool, dope, awesome, fun, good}* with 93 terms.

In hindsight, this was a very poor choice of gold standard. The reasons will be revealed in the error analysis. However, I

believe that the reasons for using an existing resource are well justified.

Results will be presenting with this gold standard, and then revised results with a new gold standard will be presented to show the effect of human intervention.

6 Implementation

The implementation of the actual similarity measures can be divided into four steps: corpus filtering, pre-processing, similarity measures and synset assignment. The similarity measures implemented are LESK (Kilgarriff & Rosenzweig, 2000), Corpus LESK (Vasilescu, Langlais, & Lapalme, 2004), a WordNet similarity/overlap hybrid that the author developed based on work from (Seco, Veale, & Hayes, 2004), and T-Test Distributional Similarity (Curran, 2003).

6.1 Corpus Filtering

To filter the corpus of 627 138 words into a more manageable and useful size, we first removed any definitions that we did not have a gold standard entry for. This reduced the corpus to about 4000 definitions. Next, we only chose the first and most popular definition for each term in order to increase the likelihood that we have the most accurate sense for a word (this is not always the case, the error analysis section indicates that single definitions often contain multiple senses). This reduced the corpus to about 1250 definitions. Finally, we removed any definitions that did not have at least twice as many *likes* as *dislikes*, to ensure that we only had high quality definitions from the point of view of the community. This reduced the corpus to 1074 definitions.

6.2 Pre-Processing

For all entries in the corpus, some pre-processing was performed. First, all punctuation was removed. Next, any self-reference was removed, were self-reference means that the definition contains the word that it is trying to define. Just the word was removed. Next, stop words were removed using a stop list. Finally, Porter stemming (Porter, 1980) was performed on all of the words in the corpus except for the terms being defined. Stemming was not conducted only for the WordNet method of similarity assessment. The definition and example were used as signatures for all of the similarity assessments.

6.3 Similarity Measures

LESK

LESK (Kilgarriff & Rosenzweig, 2000) measures the number of words that overlap between signatures. Due to some issues with verbosity, this measure was normalized by the product of the lengths of compared signatures in number of words. This is a similarity measure; therefore the larger normalized overlap means that two words are more similar.

Corpus LESK

Corpus LESK (Vasilescu, Langlais, & Lapalme, 2004) is the same as LESK, except the overlap of words is weighted by their inverse document frequency (IDF). In this work, a signa-

ture for a term counts as a document. Therefore, words that appear in many signatures are given less weight.

WordNet Similarity/Overlap Hybrid

This measure does not use the overlap of words, but rather, their WordNet similarity, as developed by (Seco, Veale, & Hayes, 2004), related to Resnik similarity (Resnik, 1995) which measures the information content of the lowest common subsumer of two nodes using the hyponym relationship of words. Similarity is measured using the following equation, where values are, once again, normalized for length:

$$Overlap(a, b) = \frac{\sum_{w \in Sign.(a)} \sum_{x \in Sign.(b)} WordNetSim(w, x)}{Length(Sign.(a)) * Length(Sign.(b))}$$

Equation 1. WordNet similarity of two terms.

If words are not found in WordNet, then LESK overlap applies, where overlapping words are given a WordNet Similarity of 1. WordNetSim in Equation 1 gives a value between 0 and 1.

T-Test Distributional Similarity

Finally, a distributional method was implemented. For this method, we compute feature vectors for each term in the corpus and measure similarity by the cosine between feature vectors. The features in this case are the t-test association measure (Curran, 2003), as defined by the following equation:

$$ttest_{assoc}(w, f) = \frac{P(w, f) - P(w)P(f)}{\sqrt{P(f)P(w)}}$$

Equation 2. T-Test association measure.

Where f is the term, and w is a word in a signature. w, f is measured over the co-occurrence window specified by the signature of f . The term f has a feature value for each w in its signature, or is 0 if w is not within f 's signature.

This is slightly different from standard distributional methods since it does not look in a fixed window around a word as a context. Since we are only using one definition and one example for a word, there are not enough examples of use to establish a context. For this reason, we use the signature with the assumption that similar words will have similar signatures.

6.4 Synset Assignment

Assignment of synsets was conducted using a 'leave-one-out' strategy, where the synset of the target term is assigned as the true synset of the most similar word found.

7 Results

Using each of the similarity methods explained and the 'leave-one-out' strategy described above, each term in the corpus was assigned to a synset. The following accuracies were attained using each similarity measure:

Table 1. Results of synset assignment using the original corpus and gold standard.

Similarity Measure	Accuracy
Most Probable Synset	5.3%
LESK	8.0%
Corpus LESK	8.1%
WordNet Hybrid	N/A
T-Test Distributional	9.1%

As one can see in the table above, synset assignment accuracy based on the assignment of 1074 terms into 353 synsets is extremely poor. Given a 'leave-one-out' strategy, this level of accuracy would be unacceptable for synset building or WordNet augmentation.

Note that the WordNet Hybrid was not used since the implementation makes an HTTP request for every similarity assessment made. The running time is measured in hours, and this speed is impractical for this size of corpus. This is, however, not a limitation of the method, just the implementation. With local storage, running time would complete in a matter of seconds.

8 Error Analysis

Due to the exceptionally poor performance of conventional similarity methods using the corpus and gold standard, we decided to examine the entities involved to point out any sources of error, or properties that would be problematic for conventional computational techniques. This would give a good idea of the problems that would have to be solved in order for this corpus to be viable for use in semantic similarity.

The Urban Dictionary had nine properties that would be problematic for computational techniques. These included redundancy, self-reference, spelling/grammar, intra-colloquia, opinions, contrast, verbosity, multiple definitions, and no parts of speech. Each of these issues will be presented, with examples and an explanation of why they will be problematic for conventional techniques.

The gold standard also had its problems. Given modern slang, it was extremely inaccurate, and should not have been chosen as a gold standard. Nevertheless, it was an existing resource, therefore assessing its usefulness is an important undertaking. The gold standard's problems will also be addressed in this section.

8.1 Redundancy

Most terms in The Urban Dictionary have multiple definitions for the same sense due to the un-moderated nature of the system.

Term: *awesome* (3)

Definition 1: *formidable, amazing, heart-stirring, wonderful*

Definition 2: *cool, hip, exciting*

All of the definitions in (3) denote a sense of positive evaluation. They give the same information, and yet they are completely separate entities. Since senses are not annotated, if one

wanted to make use of all of the definitions in The Urban Dictionary, they would have to perform sense disambiguation within their technique; this is essentially the problem of semantic similarity. We avoided this issue by only using only one definition. However, this does not address the problem of slang ambiguity, as alternate, valid senses are being discarded.

8.2 Self-Reference

Given the nature of The Urban Dictionary as a community, there are meta-comments embedded within definitions, and many definitions contain the term in question, which provides no new semantic information.

Term: *gnarly* (4)

Definition: *These kooks have no idea what GNARLY is...*

These meta-comments have no semantic value for the sense of the term, but rather as an evaluation of other definitions. If one could identify these meta-comments, they could be leveraged for assessing controversiality or uncertainty. We simply removed any terms in the definitions that were equivalent to the target term, but could not identify meta-comments.

8.3 Spelling/Grammar

Spelling and grammar was fairly poor throughout user contributed definitions. This is fairly standard for internet websites, as humans are easily able to parse and understand poorly spelled or ungrammatical sentences.

However, this is not the case for computational techniques. Packages such as MINIPAR (Lin D., 1999), which is a dependency tree parser, was developed and tested using grammatical sentences; therefore it does not work well for ill-formed sentences (Rahman, et al., 2002). Also, part of speech tagging must be conducted with n-gram methods, as tree parsers will often not find a proper structure.

Poor spelling causes issues since single lexemes may be represented by many orthographic forms. These orthographic forms must be resolved to their lexemes in order to make use of the words as useful semantic content; otherwise, misspellings are simply low probability words.

8.4 Intra-Colloquia

Many definitions contain a great deal of slang to describe a term. This is not surprising, as the people that define the dictionary are also slang users, therefore they will use their preferred language to describe words.

Term: *hobosexual* (5)

Definition: *The opposite of metrosexual.*

The above definition uses the term *metrosexual*, which is a term that describes someone who dresses very nicely. Unfortunately, this is problematic for computational methods since these the meaning of these words is unknown due to the current poor understanding of slang; only their orthographic form can be used as information. This also causes poor overlap between definitions, as definitions may be using synonymous words, but this fact is unknown. One would have to resolve

this issue by having a high quality representation of slang, and then using this to bootstrap any computational techniques.

8.5 Opinions

Many definitions contain opinions about a term, rather than any useful semantic content. These definitions are often highly rated since many users will agree with the assertions made within the definition.

Term: *gun* (6)

Definition: *A tool that is only deadly if used for its intended purpose.*

This definition contains very little useful semantic information. The words *tool* and *deadly* has some relation to the actual term, but most of the definition is simply misleading (*intended purpose* might skew any assessments of similarity). Given the redundancy of The Urban Dictionary, one might take advantage of this to discover outlier definitions. These outliers can be tagged as not useful and subsequently ignored. The issue of outlying definitions was not addressed in this work.

8.6 Contrast

The Urban Dictionary contains definitions where definitions of antonyms are provided to give contrast to the actual meaning of the term. The issue is compounded by the fact that this contrast might be presented in novel, non-rhetorical ways; therefore it might be difficult to recognize.

Term: *punk* (7)

Definition: *TRUE: louder, faster form of rock and roll, often antiestablishment FALSE: fast, tonedead pop-rock, often about relationships*

This issue could be resolved by actually recognizing contrast, although it is difficult in some cases. Contrasting discourse connectives could be recognized in order to indicate where a definition is diverging from its actual meaning.

8.7 Verbosity

Some words in The Urban Dictionary have a great deal of pop-culture surrounding them, therefore their definitions are filled with long explanations about the word's origin, or fun facts about the word. For instance, one of the definitions for the word *poo* is 1,297 words long. This definition contains very little information about the sense of the word, rather, it contains a great deal of information about a taxonomy for the term. This taxonomy discusses issues from Satanism to ice cream. This information is not very useful for computing semantic similarity and the verbosity might inflate computations of overlap. We control for verbosity by normalizing by the product of the lengths of the definitions in words. Off-topic descriptions are not controlled for, although they might be recognized by methods similar to those suggested for opinions.

8.8 Multiple Definitions

Due to The Urban Dictionary's lack of sense annotation, many of the definitions attempt to provide this faculty by enumerating every sense of the word in a single definition. This is problematic for highly ambiguous words since the definition will have words from all over semantic space and the resulting meaning will be somewhere in the middle of all of the senses for the word. For instance, the first definition of the work *fuck* has definitions for eight senses of the word (definitions contain up to 19), with widely varying accuracy. One of the senses states: *to procreate*, where it would be better described as *sexual intercourse*, as this term has little to do with conception. Luckily, multiple definitions might be recognized since they are generally demarcated by numerals, but this was not done for our work in order to simplify the collection of the corpus, as many types of demarcation are possible.

8.9 No Parts of Speech

The Urban Dictionary does not have parts of speech listed for words. Parts of speech would aid greatly in disambiguating the sense of definitions. They would also be useful as a resource for part of speech taggers. If one wanted to determine the part of speech of a word, it would be best to conduct part of speech tagging on the examples, where the word is actually used in context (Beware: the examples are subject to all of the problems described). Parts of speech were ignored for this work as only a single definition of a word was used.

8.10 Poor Choice of Gold Standard

After examining some of the classifications made by the similarity assessors, we noticed that the gold standard was not very accurate representation of slang, or did not fit the meaning of 'synsets' very well. Here are some examples of 'misclassified' terms:

Term: *kill a kitten* (8)
Classified as: *censored replacements of offensive words*

The term *kill a kitten* is a euphemism for masturbation, therefore it does indeed fit the classified category.

Term: *skid marks* (9)
Classified as: *related to driving*

Although the above classification does not fit the slang usage of *skid marks*, it does fit the conventional usage. The Urban Dictionary has evolved to include some non-slang definitions, usually appearing along with the slang definition. A good gold standard will allow for ambiguity that crosses into non-slang.

Term: *peace* (10)
Classified as: *list of goodbyes*

This is a very obvious correct classification since *peace* is a widely used synonym for *goodbye*. The gold standard did not have *peace* listed under any synset, which brings into question the use of related words as synsets. All words have at least one sense, therefore if some words are listed without sense, then the gold standard is obviously inaccurate.

There are many more similarly correct misclassifications, but these examples are sufficient to demonstrate that the gold standard was a poor choice for the task. To have a more faithful assessment of The Urban Dictionary as a resource for slang, we would have to devise a more accurate gold standard. This is described in the following section.

9 Revised Gold Standard

To create the new gold standard, the author chose nine synsets, and assigned terms to these synsets from the set of 1074 definitions. This resulted in a corpus of 85 terms. It is not ideal for the author to do all of the annotation, and deadlines prevented the author from doing any user studies to assess semantic similarity (Miller & Charles, 1991), therefore they sent the reduced list of terms with definitions and the list of synsets to a family member, who has no experience in computational linguistics, for assignment (the author's assignments were not given). The corpus' definitions were provided to resolve any ambiguities that might exist for the sense of any terms in the corpus. Both annotators had 100% agreement, which indicates that the synsets were unique and separated enough to expect decent results with a good corpus.

The nine new synsets were, with frequencies noted in brackets: Drunk (6), Attractive (7), Sex (12), Fail (3), Significant Other (Male) (3), Significant Other (Female) (7), Positive Evaluation (29), Negative Evaluation (10), Money (8).

10 Revised Results

Using the same similarity measures and the same 'leave-one-out' strategy for the revised gold standard and reduced corpus, the following accuracies were attained:

Table 2. Results of synset assignment using the revised corpus and gold standard.

Similarity Measure	Accuracy
Most Probable Synset	34%
LESK	42%
Corpus LESK	46%
WordNet Hybrid	37%
T-Test Distributional	57%

With the new synsets and reduced corpus, accuracy is much better, although it is still unacceptable for any practical thesaurus building or WordNet augmentation.

If we examine the accuracy within each synset, a stronger argument for poor performance is attained.

As you can see in the table below, accuracy is extremely low in some of the more high frequency synsets, such as Negative Evaluation. Further exploration shows that most of the misclassified Negative Evaluations are classified as a Positive Evaluation, which is problematic for use in evaluative domains. The reason for this is likely due to the fact that negative and positive evaluations often appear in the same context. For example, *I hate you* and *I love you* are polar opposites, yet share the same context. Since Positive Evaluation is the more probable of the two, the shared context is assigned as a Positive Evaluation.

Table 3. Within-synset assignment accuracy using T-Test Distributional Similarity.

Similarity Measure	Accuracy
Money	62%
Drunk	50%
Significant Other (Male)	67%
Significant Other (Female)	0%
Sex	33%
Attractive	71%
Fail	0%
Positive Evaluation	79%
Negative Evaluation	0%

A good system would show fairly high accuracies across all synsets, as that would indicate good precision and recall. This is not the case, therefore the methods employed do not work very well given the revised gold standard and reduced corpus.

11 Lessons Learned

The primary lesson learned from doing this work is that language use on the internet is anything but conventional, which means that computational techniques that were designed for conventional language are bound to perform poorly. This fact raises issues for the prospect of natural language understanding in the internet’s more unconventional and un-moderated, but popular, communities such as MySpace, Facebook, and YouTube. These quickly evolving communities bring neologisms to life on a daily basis. To have an accurate representation of language, including slang, we must understand this use of language and develop domain-general techniques that will work in this context.

Not everyone speaks like the Wall Street Journalists. This is very true on the internet and in non-textual domains such as conversational speech. This is especially apparent with youth, who are usually at the forefront of using new technology. To develop conversational agents that can actually interact with real people using real conversation, we must not use idealized versions of language. Language is organic, ever-evolving, and does not exist without knowledge. In order for computational linguistics to be successful at modeling language for use in intelligent systems, we must not ignore these facts. Techniques need to be knowledge-driven and general in order to cope with the ever-evolving ways of human communities.

The evolution of language also indicates that systems based on statistical techniques may have a shorter lifespan in the internet domain. Learning must be continuous and online to cope with changing uses of language.

The idea of coming up with good similarity assessments of slang using this corpus was rather ambitious. To actually model slang, much more work must be done, such as some of the suggestions previously mentioned. This realization led to the learning of another lesson: to think big, but tackle problems in small parts. Research questions are often very general and can take an entire lifetime to answer. The author believes that a good strategy to conducting research is to keep the general, idealized research question in mind, while attacking the smaller problems that will eventually lead to an answer. A question that is generally unaddressed by current research will not be answered in one paper. However, understanding the problem and identifying the smaller research questions behind

it is a useful and enlightening exercise; this turned out to be the aim of this paper.

Another lesson learned was to be more careful when making result-changing decisions. The choice of the original gold standard was quite poor, and this fact ended up being the cause of a great deal of wasted effort. In the future, it will help to identify problems with a corpus or gold standard before any development is conducted. That way research can be more focused and problem-driven, rather than performing triage during development and testing.

12 Evaluation

This work is both unsuccessful and successful. It was unsuccessful at attaining the original research goal of assessing semantic similarity in the domain of slang, although this goal is, in hindsight, much more ambitious than originally conceived. The project is successful since it identified some very important research questions and issues that arise when dealing with user-contributed, un-moderated, online material. The Urban Dictionary is not suitable for use with conventional dictionary-based semantic similarity methods. It is, however, an interesting, organic example of how language is used and described on the internet. Perhaps this understanding the meaning of words through The Urban Dictionary is an ‘AI complete’ (Shapiro, 1992) problem, but it serves as an indicator of the limitations of current techniques. Language use cannot be idealized; truly successful techniques must function despite the idiosyncrasies that arise due to human intervention. This work also successfully points out some of the problems that need to be addressed with user-contributed internet dictionaries as a tool for computational linguistics.

This project was also successful in identifying the type of research that I would like to conduct in the future. I would like to avoid doing work that is subject to restrictive syntactic constraints, as well as single-use statistical methods that are not flexible to change or online learning. I would rather focus on knowledge-driven language understanding, as humans are often able to infer meaning despite a lack of syntax or conventional semantics. This research direction brings up some interesting questions that are addressed in the next section.

As an example of work that provides novel and useful techniques for semantic similarity, this work is quite weak. Many of the assumptions and decisions made were quite naïve and not very enlightening to the computational linguistics community. The actual techniques used or implementation are not very useful. However, the strength of this work lies in the realization that there is an extremely large portion of human language that has not even been considered. To understand much of the internet, we need to broaden our understanding of natural language to include slang and unconventional uses of written text. This paper serves to provide direction for future research.

13 Future Work

This work and the lessons learned from it suggest some future work in the area of slang and semantic similarity.

One obvious area for future work is in creating a WordNet for slang. Someone in the linguistics community must care enough about slang and its usage to generate a modern and up-

to-date dictionary for slang with properties that are conducive to conventional word similarity techniques. This WordNet for slang might contain synsets, antonyms, synonyms and origins in order to develop an ontology for the development of slang in order to further understand how neologisms are established. If the network of slang is rich enough, one could also describe some of the more complex relationships such as hyponym and meronym relationships.

An entity such as this WordNet for slang might be leveraged to develop a better vocabulary for evaluations. In the work of evaluative summaries, user-contributed evaluations and reviews are examined to summarize the content of these reviews. Since most modern reviews appear on the internet, and internet users use slang and unconventional language, we might want to have a better understanding of what kinds of words people use to evaluate things. For instance, using the word *wicked* as a negative evaluation devised from some bootstrapping technique with WordNet would be seen as unnatural by internet users. An accurate description of slang might enhance review understanding and generation.

Given a high-quality subset of modern slang, it would be interesting to see if there are bootstrapping techniques that can augment the network automatically by examining internet communities such as The Urban Dictionary, MySpace, or YouTube. This work would involve an accurate assessment of semantic similarity and synset assignment, the original goal of this project. Naturally, this work would have to address all of the problems explained in this paper and other idiosyncrasies that might arise due to a community's use of communication tools.

The use of poor grammar by internet users suggests that language understanding might not be based in syntax. One interesting question is whether there is an underlying syntax behind what people are saying on the internet, or if understanding is purely semantic, with no underlying, statistically validated syntax. Humans have an amazing ability to effortlessly understand ill-formed sentences. Might this be a result indicative of how we actually process linguistic information?

One very interesting problem that the author is interested in pursuing is knowledge-driven vocabulary building. Humans are often able to understand the meaning of words through context without prior experience. This faculty most likely has to do with knowledge and inference; therefore, an interesting question is, given an ontology and some context, can we infer the meaning of words? This would be quite useful in internet domains, as neologisms are being invented every day and they are very rarely explained to anyone. One domain where this occurs often is in online multiplayer games. Players come up with new acronyms, verbs and nouns and they are adopted by the community effortlessly. Given that video games have a fairly well-defined ontology, this would be a good domain in which to examine the answers to these questions. To answer these questions, we might also have to examine how neologisms are invented. Are there well-defined predictors and rules to creating neologisms, or is it something unpredictable and chaotic? With the internet, we have a great deal of information about domains where neologisms arise quite quickly. How do people combine domain knowledge with linguistic knowledge to create new words?

Given that slang use is so ubiquitous, if we wish to truly understand natural language, we must understand slang and unconventional uses of language. This domain remains rela-

tively unexplored, but it is extremely important and ripe for future research; I could not have possibly covered all of the computational tasks that might be associated with understanding this use of language.

References

- Cooper, M. C. (2005). A Mathematical Model of Historical Semantics and the Grouping of World Meanings into Concepts. *Association for Computational Linguistics*, 31 (2), 227-248.
- Curran, J. (2003). From Distributional to Semantic Similarity. University of Edinburgh.
- Kilgarriff, A., & Rosenzweig, J. (2000). Framework and Results for English SENSEVAL. *Computers and the Humanities* (34), 15-48.
- Lin, D. (1998). An information-theoretic definition of similarity. *ICML 1998*, (pp. 296-304). San Francisco.
- Lin, D. (1999, March 12). MINIPAR - A minimalist parser. *Maryland Linguistics Colloquium*.
- Miller, G. A., & Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6 (1), 1-28.
- Millhauser, M. (1952). The Case Against Slang. *41* (6), 306-309.
- Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14 (3), 130-137.
- Rahman, A., Alam, H., Cheng, H., Llido, P., Tarnikova, Y., Kumar, A., et al. (2002). Fusion of two parsers for a natural language processing toolkit. *Proceedings of the Fifth International Conference on Information Fusion* (pp. 228-234). IEEE.
- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *International Joint Conference for Artificial Intelligence* (pp. 448-453). IJCAI-95.
- Seco, N., Veale, T., & Hayes, J. (2004). An Intrinsic Information Content Metric for Semantic Similarity Using WordNet. *16th European Conference on Artificial Intelligence*. Valencia, Spain: ECAI-2004.
- Shapiro, S. (1992). AI Complete Tasks. In S. C. Shapiro (Ed.), *Encyclopedia of Artificial Intelligence* (2nd Edition ed., pp. 54-57). New York: John Wiley.
- Vasilescu, F., Langlais, P., & Lapalme, G. (2004). Evaluating variants of the lesk approach for disambiguating words. *LREC-04* (pp. 633-636). Lisbon, Portugal: ELRA.