Towards Augmenting Lexical Resources for Slang and African American English

Alyssa Hwang
Computer Science Department
Columbia University
a.hwang@columbia.edu

William R. Frey
School of Social Work
Columbia University
w.frey@columbia.edu

Kathleen McKeown
Computer Science Department
Columbia University
kathy@cs.columbia.edu

Abstract

Researchers in natural language processing have developed large, robust resources for understanding formal Standard American English (SAE), but we lack similar resources for variations of English, such as slang and African American English (AAE). In this work, we use word embeddings and clustering algorithms to group semantically similar words in three datasets, two of which contain high incidence of slang and AAE. Since high-quality clusters would contain related words, we could also infer the meaning of an unfamiliar word based on the meanings of words clustered with it. After clustering, we compute precision and recall scores using WordNet and ConceptNet as gold standards and show that these scores are unimportant when the given resources do not fully represent slang and AAE. Amazon Mechanical Turk and expert evaluations show that clusters with low precision can still be considered high quality, and we propose the new Cluster Split Score as a metric for machine-generated clusters. These contributions emphasize the gap in natural language processing research for variations of English and motivate further work to close it.

1 Introduction

Current research in natural language processing has a fundamental gap: we lack strong resources for variations of English other than formal Standard American English (SAE), including slang and other dialects. African American English (AAE) and slang, in particular, are both very common genres of modern English, especially on social media. This lack of lexical resources prevents us from creating technology for analyzing dialects like these, which evolve quickly on social media. We already have lexical resources for understanding SAE (e.g., the Dictionary of Affect in Language for sentiment, WordNet for synonyms and antonyms, and ConceptNet for a variety of word relations), but equivalent resources for nonstandard English do not exist (Whissell, 1989; University, 2010; Speer et al., 2018). The small WordNet-like resource for slang, SlangNet, contains only 3000 words (compared to the 150000+ words in WordNet), and there are no resources for AAE at the time of this study (Dhuliawala et al., 2016). To develop tools for analyzing nonstandard English, we need lexical resources that can provide clues about the meaning of new words that appear. We also need approaches for evaluating whether the derived representations are accurate or not.

In this paper, we present a comparison of methods that combine clustering algorithms and word embeddings to group unknown words in the semantic space. We explore the use of agglomerative and k-means clustering on GloVe and Word2Vec embeddings and Brown's clustering on bigrams to create semantically related clusters of words which could then be used in downstream tasks. The overall goal is to create clusters of semantically related words found in two datasets with large amounts of AAE and slang, but we first conduct preliminary exploration on the smaller, more standard Brown Corpus. This corpus gives us a quick way to evaluate combinations of clustering algorithms and word embeddings, given that it uses a standard, formal dialect of English on which our algorithms can produce more directly interpretable clusters. After determining the optimal clustering algorithm and word embedding combination based on the preliminary exploration with the Brown corpus, we use the optimal combination to create clusters of semantically related words found in the experimental datasets.

After clustering words in the Brown corpus, we need an automatic way to determine if the clusters are accurate. We use WordNet synsets and ConceptNet "relatedto" relations as the gold standards for sets of semantically related words and calculate the precision and recall of machine-generated clusters, but we note that even SAE is better evaluated using human judgements. After evaluating these measures for all combinations of clustering algorithms and word embeddings on the Brown Corpus, we decide to move forward with GloVe embeddings and k-means clustering as the optimal combination on the two experimental corpora. The first corpus, TwitterAAE, was collected by the SLANG Lab at the University of Massachusetts at Amherst and slightly resembles SAE (Blodgett et al., 2016). The Gang Violence dataset, collected by the SAFE Lab at Columbia university, is the second corpus; it contains AAE combined with hyper-local slang that make the text especially difficult to understand (Blevins et al., 2016). We refer to the two AAE corpora as the experimental datasets.

By clustering unknown words with related known words, our approach can expand existing resources for automatically learning slang and AAE words. Throughout our work, we learn that established lexical resources like WordNet and ConceptNet need to be augmented to support slang and AAE: these resources are unable to serve as gold standards for words that have never been seen before. We also provide results from human evaluations on subsets of our machine-generated clusters. An evaluation on Amazon Mechanical Turk (AMT) over the clusters produced by the Brown Corpus shows that our machine-generated clusters are of high quality even when precision and recall scores may disagree. An expert from the Columbia University School of Social Work also evaluated a small set of example clusters from the experimental datasets because the Turkers would not be familiar with the slang in these corpora. Our efforts in combining clustering algorithms with word embeddings, automatically calculating precision and recall based on established lexical resources as gold standards, and verifying our results with human annotators show that we have made much progress towards methods for inferring the semantic meaning of slang and AAE, but much work remains to expand lexical resources for these variations of English.

This work makes the following contributions for analyzing slang and AAE:

- Methods for representing new words by clustering them with semantically related known words, which will ultimately help in augmenting lexical resources for slang and AAE,
- Exploration of clustering algorithms and word embeddings for clustering in the semantic space, and
- Automatic and human evaluation of these machine-generated clusters from three different datasets, along with the Cluster Split Score as a new metric for evaluating the clusters.

2 Related Work

Researchers have already started to work on applications of natural language processing for social media and nonstandard English. One approach includes adapting pretrained word embeddings for target domains, like social media. Recently, Chang et al. (2018) generated a lexicon and set of domain-specific word embeddings that were automatically induced from an unlabeled section of the Gang Violence dataset that is used in this work. Han and Eisenstein (2019)'s work on fine-tuning BERT embeddings for Early Modern English and Twitter supports the viability of using domain-adaptive fine-tuning for social media. Costa Bertaglia and Volpe Nunes (2016) propose an unsupervised, scalable, and language- and domain-independent method for learning word embeddings for Brazilian Portuguese. Other unsupervised methods include Hamilton et al. (2016)'s label propagation framework to induce domain-specific sentiment lexicons using seed words and Sinha and Mihalcea (2007)'s graph-based word sense disambiguation. Fine-tuning pretrained word embeddings is a promising foundation for developing completely unsupervised algorithms for semantics. In our work, we focus on simple methods using word embeddings to create clusters of semantically related words, an approach that enables interpretability of results as well as information about the meaning of new words that are just beginning to appear with low frequency.

Advancements in NLP methods related to nonstandard English could inform research in sociolinguistics and dialectology, which typically use other methods (Meyerhoff, 2016). Timestamps on social media

datasets also allow for observing and predicting the evolution of language. Robust systems for tracking the appearance of new words, association of new meanings with existing words, and disappearance of old words can help us understand how, when, and why language changes. Eisenstein et al. (2014) show that language evolution in computer-mediated communication reflects geographic proximity, population size, and racial demographics. Stewart and Eisenstein (2018) find that linguistic dissemination is a strong predictor of the longevity of a new word while social dissemination is not. Change in online language is driven by social dynamics and sociocultural influence, and using natural language processing techniques on large social media datasets yields important results for sociological studies (Goel et al., 2016). Along with tracking the evolution of language, our methods can help with the use of word embeddings for lexical discovery (Roberts and Egg, 2018). Our work aims to embrace the constantly evolving nature of language and improve the representation of meanings of new words introduced over time.

3 Corpora and Linguistic Tools

This paper presents clusters of vocabulary from three corpora: Brown Corpus, TwitterAAE, and the Gang Violence dataset. We also use WordNet and ConceptNet to automatically calculate precision and recall of automatically generated clusters. We present example sentences from each corpus below.

Brown Corpus	"He's all right, Craig," Rachel said.
TwitterAAE	Whoever tryna do this tax thing to get more bread let me know
Gang Violence	people just .i.p dis [emoji] [emoji] i

3.1 Brown Corpus

The Brown Corpus is a collection of formal SAE sources printed in the mid-1900s (Francis and Kucera, 1961). The full corpus contains a million words, but we use the fiction subsection, which contains 7,000 words, for preliminary analysis. This smaller dataset mostly contains words that are widely familiar in SAE, making the results more easily interpretable. For this reason, we use the Brown Corpus as a preliminary step to test the algorithms and automatically interpret the results.

3.2 TwitterAAE

The SLANG Lab at the University of Massachusetts at Amherst has developed a corpus of 830,000 tweets (500,000 words) aligned with African American demographics. This work is provided as an extension of previous work done to identify tweets written in AAE based on geo-location and similarity to a harvested sample of tweets verified to be written in AAE (Blodgett et al., 2016). The language in these tweets is more similar to SAE than the language in the Gang Violence dataset and contains less slang, making it an ideal experimental dataset to bridge the gap between SAE and AAE.

3.3 Gang Violence

The Gang Violence dataset is a collection of 5,000 labeled tweets written by Gakirah Barnes, a deceased member of a Chicago gang, and her top communicators on Twitter (Chang et al., 2018). This corpus is an expansion of a previous corpus collected by natural language processing researchers at Columbia University in collaboration with the School of Social Work, and research assistants from Chicago neighborhoods with high rates of violence confirmed that much of the language in the corpus differed from SAE (Blevins et al., 2016). Along with giving different meanings to words already seen in SAE resources (such as "ion" as an abbreviation for "I don't" instead of a type of atom), the authors of the tweets create new words at a rapid pace to describe recently occurring events. The high rate of unknown words and familiar dialect of English make this corpus a challenging one to parse, but it will also serve as an interesting extension of TwitterAAE.

3.4 WordNet and ConceptNet

WordNet is a lexical database for English nouns, verbs, adjectives, and adverbs, which are grouped into synsets for each sense of the word (University, 2010). Synsets are composed of synonyms, or words that denote the same concept and are interchangeable in different contexts. Synsets are linked by a

small number of conceptual relations. ConceptNet, on the other hand, is a crowd-sourced multilingual knowledge graph that expands on conceptual relations (Speer et al., 2018). We use the "relatedto" relation to construct the gold cluster of semantically similar words; because "relatedto" is a more general relation than synonym, the ConceptNet clusters tend to be quite large. These two tools help us automatically generate precision and recall scores of the clusters we create. See Table 1 for a comparison of sample clusters from both resources.

4 Methods

First, we evaluate three clustering algorithms (Brown's, agglomerative, and k-means) and two word embeddings (GloVe and Word2Vec) on the smaller Brown Corpus for preliminary analysis. We then use the algorithm and embedding (k-means with Word2Vec) with the best preliminary performance to cluster the TwitterAAE and Gang Violence corpora.

4.1 Clustering Algorithms

Agglomerative clustering uses a bottom-up approach that starts with having each word in its own cluster, then pairwise combining clusters that minimize similarity distance until only one cluster containing the entire vocabulary remains (Pedregosa et al., 2011). In this case, agglomerative clustering seeks to minimize the cosine similarity between word embeddings. This recursive algorithm produces a hierarchy of clusters and allows us to examine any number of clusters from 1 to N, where N is the size of the vocabulary.

	WordNet	**Both**	**ConceptNet**
Smile	[None]	smile, grin, grinning, smiling	action, smiler + *89 words*
Blue	drab, grim, Amytal + *36 words*	blue, profane, dark + *6 words*	blow, calypso, windows + *249 words*
Sword	[None]	sword, steel, brand, blade	tuck, sword-bearing + *182 words*

Table 1: The gold clusters from WordNet and ConceptNet are shown in the table above. For a query word shown in the leftmost column, the synonyms in only WordNet, related words in only ConceptNet, and words in both resources are displayed in the next columns.

Brown's clustering is a type of agglomerative clustering that uses context to group similar words together (Brown et al., 1992). In Brown's clustering, we use bigrams to account for context and pairwise combine clusters whose words share similar neighbors; Brown's clustering groups individual words together based on context from bigrams. This clustering algorithm can be used to assign words to classes based on the clustering results, which would allow for the categorization of new words in the future. A class can function as the high-level label for a cluster of words. The original work presents classes and clusters built from a 260,000-word vocabulary, such as:

> Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends ...
>
> mother wife father son husband brother daughter sister boss uncle
>
> feet miles pounds degrees inches barrels tons acres meters bytes

Brown's clustering algorithm was able to group words of similar class: days of the week, family members, and units of measure, from the example taken from the original work above. It was able to group misspellings: { *that, tha, theat* } were clustered together as typos for *that*. It also accounts for "sticky pairs," pairs of words that are found in a specific order more often than alone or in reverse order, like *Humpty Dumpty*, *Ku Klux*, *Klux Klan*, and *mumbo jumbo*. These results seem promising, but the algorithm was trained on longer, more formal sources of SAE long before the rise of social media. Lack of context, especially in short tweets, may pose an issue for Brown's clustering, but this clustering algorithm accepts raw text as input rather than vectorized representations of words. This eliminates one step of preprocessing and makes it a helpful preliminary experiment.

K-means clustering partitions its input into k groups by randomly initiating cluster centers and determining which words are closest to those centers (Bird et al., 2009). The centers are then set to be the mean of all the data points that belong to the cluster. These two steps are repeated for a set number of iterations or until some level of stability is achieved. This clustering algorithm is well-documented, simple to implement, and scalable to large projects, but it requires choosing k manually and struggles with clusters of varying sizes. Given the variety of language in the corpora, the clusters are not likely to be of uniform size and choosing the number of clusters beforehand requires additional domain knowledge.

4.2 Word Embeddings

Word2Vec is a neural model trained on Google News (Mikolov et al., 2013). The older of the two word embeddings, Word2Vec represents a baseline for word embedding results. Like GloVe, Word2Vec is context-independent and combines all senses of a word into a single vector.

GloVe embeddings are similar to Word2Vec, but they are trained on a cooccurrence matrix rather than a neural network (Pennington et al., 2014). Using a context-independent word embedding loses the distinction between different senses of a word but makes the vectors immediately available for downstream tasks. We use the 50-dimensional Twitter embeddings for this task.

BERT embeddings are the current state of the art, but we do not use them for this work. These embeddings are trained on a context-dependent neural model at the subword level, which makes BERT more robust to out-of-vocabulary words and would be useful for the constantly changing language on social media (Devlin et al., 2019). BERT embeddings, however, separate the different senses for each word, so they cannot be used without the model. The size of the BERT model and embeddings combined with the number of words for each corpus created an extremely high demand for memory that made BERT a poor candidate for this particular clustering task.

5 Results and Discussion

5.1 Preliminary Analysis: Brown Corpus

The first task for this work was to try all of the clustering algorithms and word embeddings on the smaller Brown Corpus and choose the highest performing pair for the experimental datasets. Brown's clustering with bigrams for context history yielded interesting clusters like

> that, as, when, what, which, if, where

> be, have, do, get, go, see, make, take, think, tell, find, hear

but this particular clustering algorithm is incompatible with word embeddings and poorly maintained, making it a weak choice for future work.

The results for agglomerative and k-means clustering varied distinctly in cluster size distribution (see Figure 1). Agglomerative clustering produced unbalanced clusters, with over 20% of clusters containing two words and a couple containing over 900 words. K-means clustering produced much more balanced clusters, with over half the clusters containing ten or less words. Although the results for Word2Vec and GloVe embeddings were similar, we choose to continue working with GloVe because of its faster runtime and availability of Twitter embeddings. Based on these findings, the rest of the experiments are run with k-means clustering and GloVe.

5.2 Sample Clusters from Brown, TwitterAAE, and Gang Violence Datasets

K-means clustering and GloVe embeddings produce highly interpretable clusters of semantically similar words, many of which made intuitive sense. Clusters from the Brown Corpus include

1. flashes, rosy

2. apple, blackberry, camera, flash, led, messenger, notebook, opera, telephone, windows

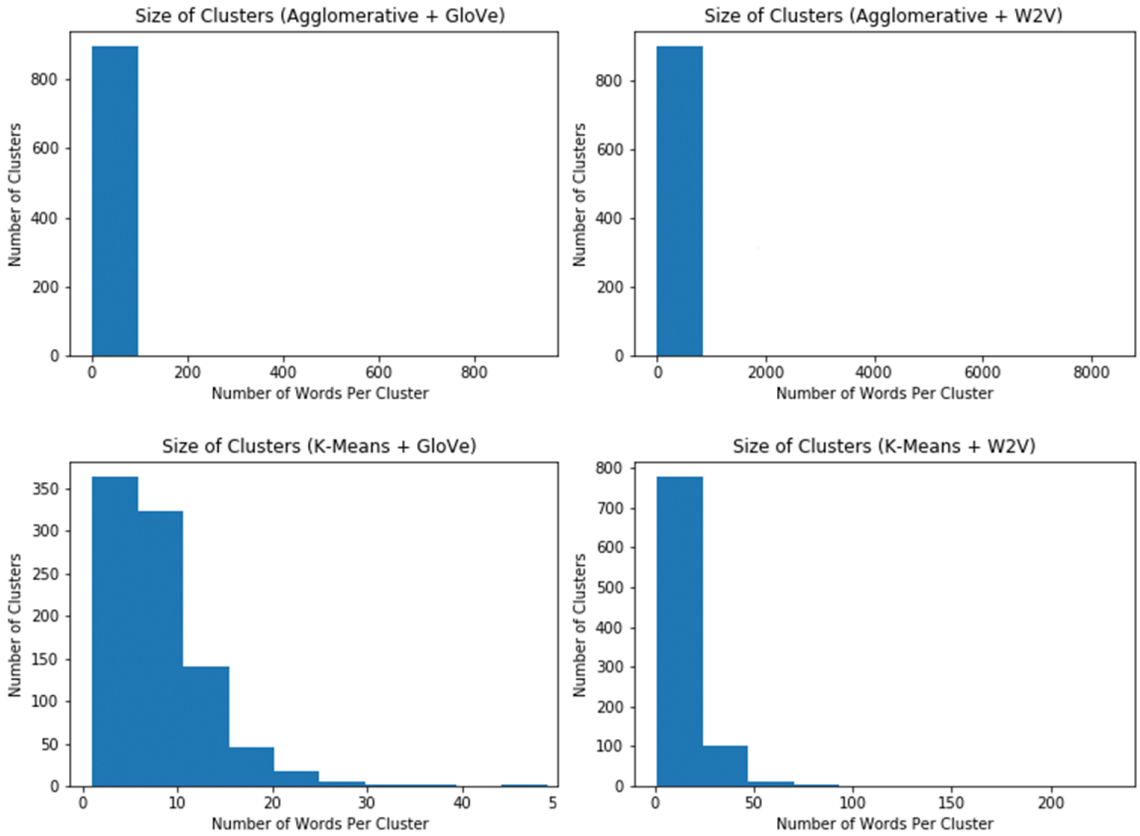

Figure 1: Sizes of word clusters for preliminary analysis on Brown Corpus with agglomerative and k-means clustering and GloVe and Word2Vec embeddings.

3. alors, chambre, corps, et, fille, fond, genre, lit, merveilleux, oui, petit, petits, plus, pour, repose, tout, week-end

Smaller clusters, like cluster 1, contain little interesting information, but clusters containing approximately ten words show obvious relationships. Cluster 2 contains words related to technology, like the Apple and Blackberry companies; Messenger, Windows, and Opera software; and telephone and camera. In this cluster, *led* may refer to an LED light or screen. Cluster 3 contains French words, including words that have meaning in both English and French like *fond* (Fr: to melt), *genre*, *lit* (Fr: bed), *plus* (Fr: more), *pour* (Fr: for), *repose* (Fr: to relax), and *week-end*. These results may seem disappointing, but this work was in English only and we did not use cross-lingual embeddings. Therefore, we were unable to compute semantic similarity across languages. These French words represent a very small portion of the dataset, which were clustered together as misfits that do not belong in other clusters, so it makes sense that they are seen together here.

Because of memory constraints related to the size of the TwitterAAE corpus (about 550,000 words), the clustering algorithm has to be run in sections. This generates many singleton and two-word clusters, and related words may not have been clustered together if they were in separate sections. Sample clusters are shown below:

4. ads, content, pinterest
5. pinger, iphonee, goom
6. cults, exorcisms, croak
7. lieing, talmbout, trippen

The words in cluster 4 are also familiar in SAE: *Pinterest* is a website that displays creative *content* and *ads*. Cluster 6 is also composed of familiar words that are related to each other. *Cults* are sometimes portrayed as performing *exorcisms*; the word *croak* may not make sense in this context, but it can also

be a slang term for death. Cluster 5 contains mostly slang words, like a misspelling of iPhone and *pinger*, which may refer to a phone that receives "pings." The word *goom* is completely unfamiliar, but the clustering implies that it is related to cell phones (reliably determining if this is true is why we still depend on human annotation). Cluster 7 contains even more unfamiliar slang words, like *lieing* (a misspelling of lying), *talmbout* (a shortening of "talking about"), and *trippen* (an alternative spelling for "tripping," which is slang for behaving wildly). These words are related in context but show that the interpretability of these clusters requires domain knowledge on the part of the evaluator.

The individual words in the Gang Violence dataset are very unfamiliar but the clusters produced by the k-means algorithm and GloVe embeddings lend more insight to the meanings of the words:

8. chat, dm, fb, inbox, insta, offline, skype
9. app, etc, facebook, google, instagram, internet, mail, twitter, whatsapp, wifi
10. ils, ont, qui, sont
11. beto, cris, diego, felipe, fran, gabriel, lucas, manu, pedro, rafa, santos, victor
12. bla, cuba, keta, kt, mcm, mmg, mula, ni, nk, pon, pun, tk

Clusters 8 and 9 contain words that are related to social media, but cluster 8 appears to be more slang while cluster 9 contains more formal language. The words *facebook* and *instagram*, for example, appear in cluster 9 while cluster 8 contains the popular abbreviations for the two social networking sites. Cluster 10 contains exclusively French words, a repeat pattern from cluster 3 of the Brown Corpus. This may be a coincidence or a testament to the GloVe pretraining algorithm, but these words are also grammatically related: *ils* is the male plural pronoun for they, and the verbs *ont* and *sont* are the conjugations of avoir (to have) and être (to be) for *ils*. Cluster 11 appears to be a collection of men's first names of Hispanic origin. Cluster 12 is another, more powerful example of the importance of domain knowledge for clustering slang in the semantic space. These words may be slang, abbreviations, or a different language, but it is difficult to tell; current resources like WordNet and ConceptNet are also unlikely to contain much information.

5.3 Precision, Recall, and Unknown Words

The precision (Equation 1) and recall (Equation 2) for each word and cluster are calculated based on WordNet and ConceptNet as the gold standards (see Tables 2a and 2b).

$$\text{precision} = \frac{\text{size}\left(\text{cluster}_e \cap \text{cluster}_g\right)}{\text{size}\left(\text{cluster}_\text{e}\right)} \tag{1}$$

$$\text{recall} = \frac{\text{size}\left(\text{cluster}_e \cap \text{cluster}_g\right)}{\text{size}\left(\text{cluster}_\text{g}\right)} \tag{2}$$

Where $e =$ the experimental cluster and $g =$ the gold standard from WordNet or ConceptNet.

These scores are very low and seem to increase for the Gang Violence and TwitterAAE corpora, but this pattern is more indicative of the poor quality of WordNet and ConceptNet for evaluating variations of English than the quality of machine-generated clusters. WordNet is composed of synsets, which means that the WordNet gold cluster for a word is strictly limited to synonyms. Not all related words, however, are synonyms. The low precision scores for Brown Corpus and the Gang Violence dataset show this gap in word relations. ConceptNet expands on word relations and includes words that are "relatedto" a query word, which better captures the semantic similarity we are trying to evaluate. This gives us bigger clusters per word, which causes precision scores to drop and recall scores to rise. Some scores, however, are inflated–which appears to be the case for the experimental datasets–because a large portion of words are not seen in either of these linguistic tools (see Table 3). A word is considered semantically similar to itself, so any word that is not seen in WordNet or ConceptNet would have a gold cluster of just one word: itself. This lack of data causes recall to rise, especially for small clusters.

		WordNet	ConceptNet
Brown	w	0.138	0.246
	c	0.233	0.258
GV	w	0.287	0.561
	c	0.157	0.551
TwitterAAE	w	0.686	0.686
	c	0.933	0.687

(a) Precision values.

		WordNet	ConceptNet
Brown	w	0.172	0.131
	c	0.261	0.142
GV	w	0.281	1.0
	c	0.151	1.0
TwitterAAE	w	0.872	1.0
	c	0.933	1.0

(b) Recall values.

Table 2: Precision and Recall at the word (w) and cluster (c) levels for all datasets based on WordNet and ConceptNet as gold standards.

Because the precision and recall scores depend on WordNet and ConceptNet, and both of these linguistic tools only partially represent semantic similarity or do not contain the query word at all, these scores do not completely indicate the quality of the clusters. WordNet and ConceptNet are excellent tools for other tasks involving standard English, but they misrepresent the quality of clusters of variations of English. Human annotation can help better describe the quality of these clusters.

	Total	GloVe	WN	CN
Brown	6922	0	549 (7.9%)	146 (2.1%)
GV	227223	2	9884 (4.3%)	8492 (3.7%)
Twitter AAE	53642	0	50163 (93%)	32207 (60%)

Table 3: The total number of words in each dataset along with the number of words missing from the pretrained GloVe embeddings, WordNet (WN), and ConceptNet (CN).

6 Manual Evaluation with Amazon Mechanical Turk

To evaluate the quality of a cluster, Amazon Mechanical Turkers are given a machine-generated cluster of words and asked to split the cluster into subclusters of semantically similar words (a cluster that requires no splits would yield one subcluster: itself). Semantic similarity includes synonymy or relevance in the same context. Red and blue, for example, are not synonyms but they are relevant in the context of colors, so they would be considered semantically similar and clustered together. Turkers were given the following instructions (below) with three examples (full instructions with examples are provided in the appendix):

> You will be given a list of several words. These words will be separated with a space–there are no phrases or compound words. The goal is to group words into as few groups as possible by semantic similarity. Words may be grouped together if they have similar meaning or would make sense appearing in the same context with at least one other word in the cluster. Groups of singleton words are acceptable, but not every cluster should be singleton words.

We also define our own metric, the Cluster Split Score (see Equation 3), for evaluating Mechanical Turk annotations. The intuition is simple: a high-quality cluster would not need to be split into multiple clusters because all of the words should be similar to each other. This can be considered a top-down hierarchical approach. A high-quality cluster would have a low number of splits. Clusters with less than three words are omitted from this evaluation task and 300 clusters are randomly sampled without replacement from the remaining. The 300 clusters are divided into three batches of 100 clusters, and each batch is evaluated by three Turkers. We then take the mean number of splits as the raw score for each cluster.

We instruct the Turkers to perform the task on clusters from the Brown Corpus because these words are more familiar and the clusters are more easily interpretable. Although the task showed only slight agreement (Fleiss' kappa = 0.046, 0.066, 0.084 for each batch), the low number of mean splits indicate

that the clusters are of higher quality than the automated precision scores may imply (see Figure 2). This task would benefit from rigorously defining "semantic relatedness" and selecting knowledgeable annotators, especially for tasks including slang and AAE.

After Mechanical Turk evaluations are complete, we calculate the Cluster Split Score for each cluster:

$$CSS = \text{logistic}\left(\frac{\text{number of words}}{\text{number of splits}}\right) \text{where logistic}(x) = \frac{1}{1+e^{-x}}. \quad (3)$$

We divide the number of words by the number of splits because we would like to reward large clusters remaining intact while penalizing clusters being divided. This number then becomes the input for the logistic equation, which squashes the range of the scores between 0 and 1, with 1 indicating the highest quality; this ensures that Cluster Split Scores can be compared for clusters of different sizes and encourages smoother steps between scores. The automated precision and CSS are reported in Figure 2.

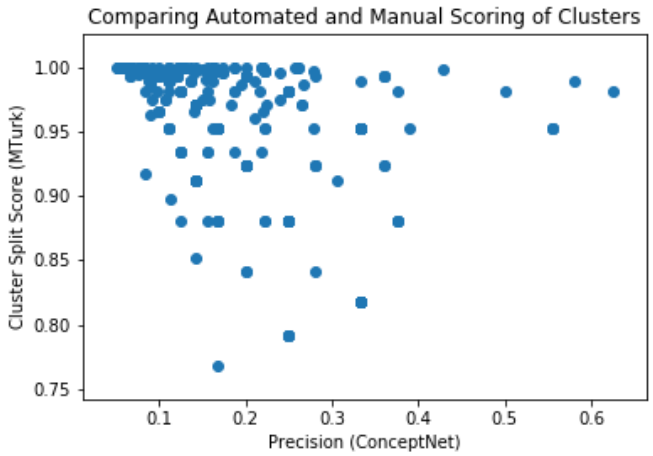

Figure 2: Automated precision scores with ConceptNet as the gold standard compared to Cluster Split Scores annotated by Mechanical Turkers.

We also ask an expert from the Columbia University School of Social Work to evaluate a small subset of clusters from both experimental datasets. Of the 14 Gang Violence clusters, he agrees with the machine generation 4 times and splits the cluster in two parts 4 times, for a mean CSS of 0.96. Of the 12 TwitterAAE clusters, he splits the cluster in two parts 12 times and never agrees with the machine-generated cluster, for a mean CSS of 0.83. The high mean Cluster Split Scores given by the expert show that machines can automatically generate high-quality clusters.

Below, we present sample clusters from the Gang Violence and TwitterAAE corpora with the expert's annotations. Each line shows one machine-generated cluster that has been split into subclusters by the expert. These clusters show many variations of English, like influences from French and Spanish and alternative spellings for existing words. Some words may be completely unfamiliar, which is why we motivate continued efforts in expanding lexical resources for slang and AAE.

Gang Violence	{ harry, mila, naya, malik, louis, payne, lou }, { hacked, dmed }, { lux }
TwitterAAE	{ shootah, bakk }, { hyz }, { bizz }
TwitterAAE	{ hungova, shleeep }, { haself }

If the automated precision scores from ConceptNet and Cluster Split Scores from AMT both reliably evaluated cluster quality, then we would see that precision and Cluster Split Scores are directly correlated. This, however, is not the case in Figure 2. The scores are skewed to the top left, showing that clusters with low precision often receive a high Cluster Split Score. This further shows that current lexical resources need to be expanded to become more inclusive of slang, AAE, and other variations of English.

7 Conclusion and Further Studies

Natural language processing resources lack representation from slang and nonstandard English that would allow us to reliably and automatically evaluate our methods. In this work, we used GloVe embeddings and k-means clustering to cluster semantically similar words in the Brown, TwitterAAE, and Gang Violence corpora. Mechanical Turk and expert observation showed promise in the quality of automatically generated clusters and that these clusters can reveal semantic relatedness between unknown words. In contrast, precision and recall scores from WordNet and ConceptNet do not currently reflect this.

We can improve this problem by expanding lexical resources for slang and AAE. Automatic clustering with simple parameters already show great potential for automatically learning the meanings of new words. This paper presents tools for automatically and manually evaluating these new resources for slang and AAE that can be created or expanded in future work. Context-dependent training methods for sense disambiguation would make clusters more robust. In addition, another important area for improvement is using domain-adaptive fine-tuning to include words that do not have pretrained embeddings and adjusting embeddings for words that may have taken on a new meaning.

Machine-generated clusters can also be used to track the evolution of language and learn new words as they appear online. This is especially important in the age of social media since new slang terms appear so quickly. Existing words can even take on new meanings; old data for existing words can then affect results if the meaning of the word has changed. The words "terrible" and "terrific," for example, used to be synonyms—our results would certainly be different if we still assumed that.

With improved word embeddings and resources for evaluation, we can build highly interpretable clusters with an algorithm as simple as k-means clustering. These tools will help natural language processing will become more inclusive of all variations of language, not just the formal, standard features that historically gathered more attention.

Acknowledgements

We would like to thank Emily Allaway, Elsbeth Turcan, and Shinya Kondo from Columbia University for their assistance. We also thank the reviewers for their time and helpful feedback.

References

S. Bird, E. Loper, and E. Klein. *Natural Language Processing with Python*. O'Reilly Media Inc, 2009.

T. Blevins, R. Kwiatkowski, J. MacBeth, K. McKeown, D. Patton, and O. Rambow. Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2196–2206, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1207.

S. L. Blodgett, L. Green, and B. O'Connor. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1120. URL https://www.aclweb.org/anthology/D16-1120.

P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-Based *n*-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–480, 1992. URL https://www.aclweb.org/anthology/J92-4003.

S. Chang, R. Zhong, E. Adams, F.-T. Lee, S. Varia, D. Patton, W. Frey, C. Kedzie, and K. McKeown. Detecting Gang-Involved Escalation on Social Media Using Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 46–56, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1005. URL https://www.aclweb.org/anthology/D18-1005.

T. F. Costa Bertaglia and M. d. G. Volpe Nunes. Exploring Word Embeddings for Unsupervised Textual User-Generated Content Normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 112–120, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/W16-3916.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for

Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

S. Dhuliawala, D. Kanojia, and P. Bhattacharyya. Slangnet: A wordnet like resource for english slang. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, page 4329–4332. European Language Resources Association (ELRA), May 2016. URL https://www.aclweb.org/anthology/L16-1686.

J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. Diffusion of Lexical Change in Social Media. *PLoS ONE*, 9(11):e113114, Nov. 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0113114. URL http://arxiv.org/abs/1210.5268. arXiv: 1210.5268.

W. N. Francis and H. Kucera. Brown Corpus Manual, 1961. URL http://icame.uib.no/brown/bcm.html.

R. Goel, S. Soni, N. Goyal, J. Paparrizos, H. Wallach, F. Diaz, and J. Eisenstein. The Social Dynamics of Language Change in Online Networks. *arXiv:1609.02075 [physics]*, Sept. 2016. URL http://arxiv.org/abs/1609.02075. arXiv: 1609.02075.

W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1057. URL http://aclweb.org/anthology/D16-1057.

X. Han and J. Eisenstein. Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1433. URL https://www.aclweb.org/anthology/D19-1433.

M. Meyerhoff. Methods, innovations and extensions: Reflections on half a century of methodology in social dialectology. *Journal of Sociolinguistics*, 20(4):431–452, Sep 2016. ISSN 1360-6441, 1467-9841. doi: 10.1111/josl.12195.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, Sept. 2013. URL http://arxiv.org/abs/1301.3781. arXiv: 1301.3781.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

J. Pennington, R. Socher, and C. Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL http://aclweb.org/anthology/D14-1162.

W. Roberts and M. Egg. A large automatically-acquired all-words list of multiword expressions scored for compositionality. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1046.

R. Sinha and R. Mihalcea. Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. page 7, 2007.

R. Speer, J. Chin, and C. Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *arXiv:1612.03975 [cs]*, Dec. 2018. URL http://arxiv.org/abs/1612.03975. arXiv: 1612.03975.

I. Stewart and J. Eisenstein. Making ”fetch” happen: The influence of social and linguistic context on nonstandard word growth and decline. *arXiv:1709.00345 [physics]*, Aug. 2018. URL `http://arxiv.org/abs/1709.00345`. arXiv: 1709.00345.

P. University. About WordNet, 2010.

C. Whissell. The dictionary of affect in language. 1989.

Appendix A. More Details on the Amazon Mechanical Turk Task

We provide more details on the evaluation task described in Section 6. We ran three batches of evaluations of 100 clusters each, with 3 Turkers evaluating each set of 100 clusters. We had three qualification requirements to help control for English fluency: (1) Location is US (2) Number of HITs Approved greater than or equal to 1000, and (3) HIT Approval Rate (%) for all Requesters' HITs greater than 97. We awarded $0.17 per assignment. See Figure A1 for an example of the user interface.

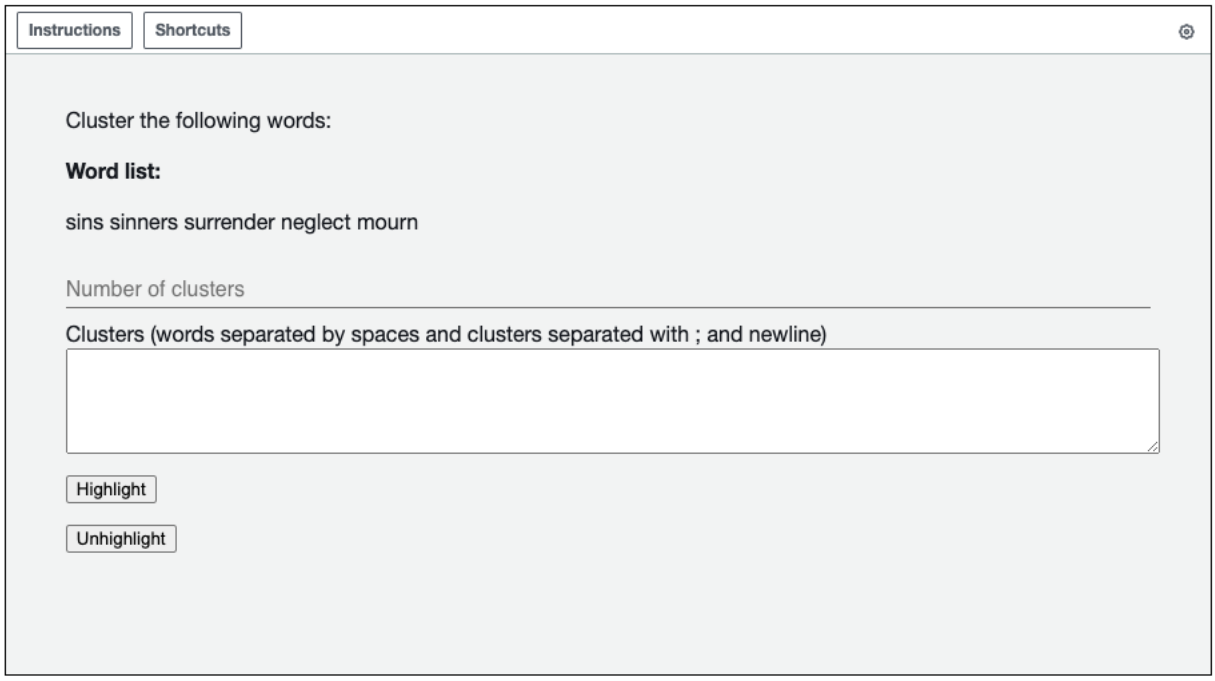

Figure A1: An example of one evaluation task on Amazon Mechanical Turk. We also included a highlight tool to help Turkers keep track of which words they have already clustered.

We provide the evaluation instructions for each task:

> Create the least number of clusters of words with similar meaning. Put a word in a cluster if at least one of its definitions is related to at least one definition of another member word. You can consider a word to be related to another if it has the same/similar meaning (like a synonym) or relate to the same topic. Sword, knife, spear, and arrow are all related to physical weapons, so they can be considered related. **You may cluster words in non-English languages with words of the same language.** You may, for example, cluster bonjour and bien together without considering the meanings of the words in the foreign language.
>
> **This task may contain foul language.**
>
> The list of words you are given will have words separated with a single space. There are no phrases or compound words. Indicate the number of clusters you make and the clusters themselves. **Separate words with a space and clusters with a semicolon (;) and newline.**

along with the examples that accompanied the instructions:

- Example 1
  - **Word List**: apple, blackberry, computer
  - **Clusters**: apple blackberry computer
  - **Number of Clusters**: 1
  - **Rationale**: Since Apple and Blackberry are technology companies and computer is a type of electronic technology, you may group all of these words in the same cluster.
- Example 2
  - **Word List**: bonjour hola oui paris
  - **Clusters**: bonjour oui paris; hola
  - **Number of Clusters**: 2
  - **Rationale**: Since this list contains French (bonjour, oui, paris) and Spanish (hola), you may split the words into two clusters.
- Example 3
  - **Word List**: blue sad red color
  - **Clusters**: blue sad red color
  - **Number of Clusters**: 1
  - **Rationale**: Blue and red are colors, but blue is also a synonmym for sadness, so the words may be grouped into one cluster.
- Example 4
  - **Word List**: shirt pants blanket book
  - **Clusters**: shirt pants; blanket; book
  - **Number of Clusters**: 3
  - **Rationale**: There should be 3 clusters (shirt, pants), (blanket), and (book). A blanket would not occur in the same context as shirt and pants (items of clothing).
- Example 5
  - **Word List**: red orange yellow green blue purple violet
  - **Clusters**: red orange yellow green blue purple violet
  - **Number of Clusters**: 1
  - **Rationale**: These words are all colors, so they can be grouped into one cluster.