Humor in Word Embeddings:
Cockamamie Gobbledegook for Nincompoops

Limor Gultchin [1]   **Genevieve Patterson** [2]   **Nancy Baym** [3]   **Nathaniel Swinger** [4]   **Adam Tauman Kalai** [3]

Abstract

While humor is often thought to be beyond the reach of Natural Language Processing, we show that several aspects of single-word humor correlate with simple linear directions in Word Embeddings. In particular: (a) the word vectors capture multiple aspects discussed in humor theories from various disciplines; (b) each individual's sense of humor can be represented by a vector, which can predict differences in people's senses of humor on new, unrated, words; and (c) upon clustering humor ratings of multiple demographic groups, different humor preferences emerge across the different groups. Humor ratings are taken from the work of Engelthaler and Hills (2017) as well as from an original crowdsourcing study of 120,000 words. Our dataset further includes annotations for the theoretically-motivated humor features we identify.

1. Introduction

Detecting and generating humor is a notoriously difficult task for AI systems. While Natural Language Processing (NLP) is making impressive advances in many frontiers such as machine translation and question answering, NLP progress on humor has been slow. This reflects the fact that humans rarely agree upon what is humorous. Multiple types of humor exist, and numerous theories were developed to explain what makes something funny. Recent research supporting the existence of single-word humor (Engelthaler & Hills, 2017; Westbury et al., 2016) defines a more manageable scope to study with existing machine learning tools. Word Embeddings (WEs) have been shown to capture numerous properties of words (e.g., Mikolov et al., 2013a;b);

coupled with single-word humor as a possible research direction, it is natural to study if and how WEs can capture this type of humor. To assess the ability of WEs to explain individual word humor, we draw on a long history of humor theories and put them to the test.

To many readers, it may not be apparent that individual words can be amusing in and of themselves, devoid of context. However, Engelthaler & Hills (2017), henceforth referred to as EH, found some words consistently rated as more humorous than others, through a crowdsourced study of about five thousand nouns. We first use their publicly available 5k mean word-humor ratings to identify a "humor vector," i.e., a linear direction, in several WEs that correlate (over 0.7) with these 5k mean humor ratings. While these correlations establish statistical significance, little insight is obtained into how the embeddings capture different aspects of humor and differences between people's senses of humor.

To complete this picture, we performed crowdsourcing studies to create additional datasets which we make publicly available: (a) beginning with a set of 120k common words and phrases chosen from a word embedding, a crowdsourcing filtering process yielded a set of 8,120 and a further set of 216 words[1] (in Appendix B) rated most humorous, (b) over 1,500 crowd workers rated these latter 216 words through six-way comparisons each yielding a personal first choice out of dozens of other personally highly-ranked words, and (c) 1,500 words (including highly-rated words drawn from these sets) were each annotated by multiple workers according to six humor features drawn from the aforementioned theories of humor.

Our analysis suggests that individual-word humor indeed possesses many aspects of humor that have been discussed in general theories of humor, and that many of these aspects of humor are captured by WEs. For example, 'incongruity theory,' which we discuss shortly, can be found in words which juxtapose surprising combinations of words

---

[1]University of Oxford [2]TRASH [3]Microsoft Research [4]Lexington High School. Correspondence to: Limor Gultchin <limor.gultchin@jesus.ox.ac.uk>.

---

[1]These words, including *gobbledegook* and *nincompoops*, were rated as more humorous than the words in the EH study, which used common words from psychology experiments. The top-rated EH words were *booty, tit, booby, hooter,* and *nitwit*.

like *hippo* and *campus* in *hippocampus*. Incongruity can also be found in words that have surprising combinations of letters or sounds (e.g. in words borrowed from foreign languages). The 'superiority theory' can be clearly seen in insulting words such as *twerp*. The 'relief theory' explains why humor may be found in taboo subjects and is found in subtle and not-so-subtle sexual and scatological connotations such as *cockamamie* and *nincompoop*. The most humorous words are found to exhibit multiple humor features, i.e., funny words are funny in multiple ways. We show that WEs capture these features to varying extents.

Further, as is understood about humor in general, single-word humor is found to be highly subjective. We show here that individual senses of humor can be well-represented by vectors. In particular, an embedding for each person's sense of humor can be approximated by averaging vectors of a handful of words they rate as funny, and these vectors generalizes to predict preferences for future unrated words. When clustered, these vector groups differ significantly by gender and age, and thus identify demographic differences in humor taste. We further introduce a "know-your-audience" test which shows that these sense-of-humor vectors can also be meaningfully differentiated using only a handful of words per person and generalize to make predictions on new unrated words.

While it would be easy to use Netflix-style collaborative filtering to predict humor ratings, WEs are shown to *generalize*: given humor ratings on some words, vector representations are able to predict mean humor ratings, humor features and differences in sense of humor on test words that have no humor ratings.

Implications. Our study suggests that some of the fundamental features of humor are computationally *simple* (e.g., linear directions in WEs). This further suggests that the difficulty in understanding humor may lie in understanding rich domains such as language. If humor is simple, then humorous sentences may be hard for computers due to inadequate sentence embeddings: if you don't speak Greek, you won't understand a simple joke in Greek. Conversely, it also suggests that as language models improve, they may naturally represent humor (or at least some types of humor) without requiring significant innovation.

Furthermore, understanding single-word humor is a possible first step in understanding humor as it occurs in language. A natural next step would be to analyze humor is phrases and then short sentences, and indeed the 120,000 tokens rated include 45,353 multi-word tokens and the set of 216 tokens rated funniest includes 41 multi-word tokens.

Organization. Section 2 defines humor features drawn from humor theories and briefly discusses related work. Section 3 describes the word embeddings and data used. Section

4 covers prediction of aggregate humor ratings from embedding vectors. Section 5 uses embedding representations to uncover group humor differences using unsupervised learning. Section 6 analyzes how word embeddings captures each of the theory-based humor features, while Section 7 analyzes how WEs capture differences in individual senses of humor.

2. Relevant Features from Humor Theories

In order to identify traces of single word humor in WEs, we first had to identify features of humor to look for. Defining humor, however, is a problem that was difficult enough to give food for thought to some of the most famous thinkers in history. Therefore, aware of our limitations, we do not seek to offer a definition of our own, but rather rely on the multiple existing theories from the social sciences, dating at least as far back as the philosophers of ancient Greece. Plato, followed by Hobbes and Descartes, all contributed to a *superiority theory* view of humor, which was formalized in the 20th century (Morreall, 2016). However, since the 18th century, two new theories of humor surfaced and became much more widely acceptable: the *relief theory* and the *incongruity theory.*

The *relief theory* offers a physiological explanation, which argues that laughter functions as a pressure relief for the nervous system, sometimes as a means to address taboo subjects that cause such stress. Proponents of this theory were Lord Shaftesbury (1709), Herbert Spencer (1911) and Sigmund Freud (1905). *Incongruity theory* explains humor as a reaction to moments of uncertainty, in which two non-related issues which seem to be unfitting are juxtaposed in one sentence or event, but are then resolved. This theory could be seen as an explanation of the intuitive importance of punch lines in common jokes, where the "set up" builds an expectation that the "punch line" violates. Albeit confusing or disorienting logically, the revelation of the incongruity creates a humorous event when this contradiction violates an expectation, in a harmless, benevolent way. Among the supporters of this method were Kant (1790), Kierkegaard (1846) and Schopenhauer (1844).

These theories are also relevant to word humor. Based on these theories and other discussions of word humor (Beard, 2009; Bown, 2015), we consider the following six features of word humor:

1. Humorous *sound* (regardless of meaning): certain words such as *bugaboo* or *razzmatazz* are funny regardless of their meaning. This feature is related to the incongruity theory in that an unusual combination of sounds that normally do not go together can make a word sound funny. This feature is also consistent with the comedy intuition of Neil Simon, Monty Python, Dr.

Seuss and the like, who discuss funny-sounding words and feature them in their works (see, e.g. Bown, 2015).

2. *Juxtaposition*/Unexpected incongruity: certain words are composed of terms which are otherwise completely unrelated. For example, there is little relation between the *hippocampus* part of the brain and the mostly un-related words *hippo* and *campus*. This feature is moti-vated by incongruity theory.

3. *Sexual* connotation: some words are explicitly sex-ual (such as *sex*) or have sexual connotations (such as *thrust reverser*). The fact that some people find those words funny can be explained by the *'relief theory'*, which suggests humor is a venting mechanism for so-cial taboos. This was also discussed in the context of computational humor by Mihalcea et al. (Mihalcea & Strapparava, 2005).

4. *Scatological* connotation: some words have connota-tions related to excrement to varying degrees, such as *nincompoop* or *apeshit*. The justification of this feature is similar to the sexual connotation feature above.

5. *Insulting* words: in the context of word humor, the superiority theory suggests that insulting words may be humorous to some people.

6. *Colloquial* words: extremely informal words such as *crapola* can be surprising and provide relief in part due to their unusually informal nature.

We study the extents to which these features correlate with humor and how well WEs capture each one. Humor is known to vary by culture and gender, and EH analyzed age and gender differences in word humor. They also found strong correlations between humor ratings and word fre-quency and length, with shorter and less frequent words tending to be rated higher. This is interesting because word length and word frequency are strongly anti-correlated (Strauss et al., 2007).

2.1. Related Work in Computational Humor

Many of these features of humor have been considered in prior work on computational humor (see, e.g. Mihalcea & Strapparava, 2005), including early work such as HA-HAcronym (Stock & Strapparava, 2003), which focused on producing humorous acronyms. More recently, Van Hee et al. (2018) focused on humor detection, Barbieri & Sag-gion (2014) utilized Twitter for the automatic detection of irony and social media was further used to explore automatic classification of types of humor, such as anecdotes, fantasy, insult, irony, etc. (Raz, 2012). Other work studies visual humor (Chandrasekaran et al., 2016) or humorous image captions (Shahaf et al., 2015; Radev et al., 2015). Other

recent work has studied satire (West & Horvitz, 2019). WEs have been used as features in a number of NLP humor sys-tems (e.g. Chen & Soo, 2018; Hossain et al., 2017; Joshi et al., 2016), though the humor inherent in individual words was not studied.

2.2. Other Work on Individual-Word Humor

In addition to EH, another psychology study found consis-tent mean humor differences in *non-word* strings (Westbury et al., 2016). Outside of research, numerous lists of inher-ently funny words compiled by comedians and authors have been published (e.g. Beard, 2009; tvtropes.org, 2017).

Finally, some comedians recommend humorous word choice. For instance, the character Willie in the Neil Si-mon play and subsequent film *The Sunshine Boys* says:

> Fifty-seven years in this business, you learn a few things. You know what words are funny and which words are not funny. Alka Seltzer is funny. You say "Alka Seltzer" you get a laugh.... Words with k in them are funny. Casey Stengel, that's a funny name. Robert Taylor is not funny. Cupcake is funny. Tomato is not funny. Cookie is funny. Cucumber is funny. Car keys. Cleveland is funny. Maryland is not funny. Then, there's chicken. Chicken is funny. Pickle is funny. All with a k. Ls are not funny. Ms are not funny. (Simon, 1974) (as cited in Kortum, 2013).

This suggests that understanding funny words in isolation may be helpful as a feature in identifying or generating longer humorous texts.

3. Data

As is common, we will abuse terminology and refer to both the words and multi-word tokens from the embeddings as words, for brevity. When we wish to distinguish one word from multiple words, we write *single words* or *phrases*.

3.1. Embeddings

WEs fit each word to a d-dimensional Euclidean vector based on text corpora of hundreds of billions of words. We consider several publicly-available pre-trained embeddings all using $d = 300$ dimensions, trained on news and web data using different algorithms. The first embedding we consider, referred to as GNEWS, is the well-studied WE trained using the word2vec algorithm on about 100 billion tokens from Google News (Mikolov et al., 2013a). The second two embeddings, referred to as WebSubword and WebFast, were both trained using the fastText algorithm on a web crawl of 600 billion tokens (Mikolov et al., 2018). The difference is

that WebSubword is trained using subword information and can therefore be evaluated on virtually any word or phrase, whereas WebFast like every other model in this paper can only be evaluated on words in its vocabulary. Finally, we consider an embedding trained on a similar web crawl of 840 billion tokens using the gloVe algorithm, referred to as WebGlove (Pennington et al., 2014).

3.2. The Engelthaler-Hill Dataset

Our first source of data is the EH dataset, which is publicly available (Engelthaler & Hills, 2017). It provides mean ratings on 4,997 single words, each rated by approximately 35 raters on a scale of 1-5. They further break down the means by gender (binary, M and F) and age (over 32 and under 32). However, since the EH data is in aggregate form, it is not possible to study questions of individual humor senses beyond age and gender.

3.3. Our Crowdsourcing Studies

The EH data serves the purpose of finding a humor direction correlating with mean humor ratings. However, in order to better understand the differences between people's preferences among the funniest of words, we sought out a smaller set of more humorous words that could be labeled by more people. Eligible words were lower-case words or phrases, i.e., strings of the twenty-six lower-case Latin letters with at most one underscore representing a space. In our study, we omitted strings that contained digits or other non-alphabetic characters.

English-speaking labelers were recruited on Amazon's Mechanical Turk platform. All labelers identified themselves as fluent in English, and 98% of them identified English as their native language. All workers were U.S.-based, except for the results in Table 4 of the appendix. We study a subset of 120,000 words and phrases from GNEWS, chosen to be the most frequent alphabetic lower-case entries from the embedding. While our study included both words and phrases, for brevity, in the tables in this paper we often present only single words. The list of 120,000 strings is included in the public dataset we are making available alongside this paper.[2]

Humor rating study. Our series of humor rating studies culminated in a set of 216 words with high mean humor ratings, which were judged by 1,678 U.S.-based raters. In each study, each participant was presented with random sets of six words and asked to select the one they found most humorous. In the first study only, the interface also enabled an individual to indicate that they found none of the six words humorous. We treat the selected word, if

any, as being labeled positive and the words not selected as being labeled negative. Prior work on rating the humor in non-words found similar results for a forced-choice design and a design in which participants rated individual words on a Likert-scale (Westbury et al., 2016). To prevent fatigue, workers were shown words in daily-limited batches of up to fifty sextuple comparisons over the course of weeks. No worker labeled more than 16 batches.

We refer to the three humor-judging studies by the numbers of words used: 120k, 8k, and 216. In the 120k study, each string was shown to at least three different participants in three different sextuples. 80,062 strings were not selected by any participant, consistent with EH's finding that the vast majority of words are not found to be funny. The 8k study applied the same procedure (except without a "none are humorous" option) to the 8,120 words that were chosen as the most humorous in a majority of the sextuples they were shown. Each word was shown to at least 15 different participants in random sextuples. To filter down to 216 words, we selected the words with the greatest fraction of positive labels. However, several near duplicate words appeared in this list. To avoid asking people to compare such similar words in the third stage, amongst words with the same stem, the word with the greatest score was selected. For instance, among the four words *wank*, *wanker*, *wankers*, and *wanking*, only *wanker* was selected for the third round as it had the greatest fraction of positive votes. The 216 words are shown in Appendix B.

In the 216 study, each participant selected not only a set of 36 "humorous" words, but further sets of 6 "more humorous" words and a single most humorous word, as follows. The 216 words were first presented randomly in 36 sextuples comprising the 216 words. In the same task, the 36 chosen words from these sextuples were shown randomly in 6 sextuples. The selected words were presented in a final sextuple from which they selected a single word. We associate a rating of 3 with the single funniest word selected at the end, a rating of 2 with the five other words shown in the final sextuple, a rating of 1 with the thirty words selected that did not receive a rating of 2 or 3, and a rating of 0 with the 180 unselected words. This process was chosen for efficiency as it requires only 43 clicks per user to identify three tiers of words that are rated highly. However, the randomness in the initial permutation of each user's words introduces noise into the ratings. Nonetheless, it is easy to see that if a participant answers consistent with a secret internal ordering π, the expected values of the ratings (0-3, as defined above) of words would also be ordered according to π.

Feature annotation study. The feature annotation study drew on the same worker pool. 1,500 words were chosen from the original 120,000 words by including the 216 words discussed above plus random words (half from the 8k study

---

[2]https://github.com/limorigu/Cockamamie-Gobbledegook

and half from the 120k). We asked participants to annotate six features discussed earlier, namely *humorous sound, juxtaposition, colloquial, insulting, sexual connotation, and scatological connotation*. Each feature was given a binary "yes/no" annotation, and results were averaged over at least 8 different raters per feature/word pair.

In each task, each rater was given the definition of a single feature and was presented with 30 words. They were asked to mark all words which possessed that feature by checking a box next to each word. A small number of workers were eliminated due to poor performance on gold-standard words. Further experimental details are in Appendix A.

4. Identifying a Mean Humor Direction

In prior work, the vector difference between embeddings of word pairs such as *trees:tree* and *she:he* was shown to be highly consistent with human ratings of concepts such as noun plurality or (binary) gender. While it is difficult to identify a word pair that represents humor, a linear direction (i.e., vector) in the embedding may still correlate strongly with mean humor ratings.

Hence, we begin by fitting a simple least-squares linear regression model, predicting the EH mean humor rating for each word in the dataset from its 300-dimensional embedding. The resulting "humor vector" is of course not humorous as a collection of 300 numbers, but its inner-product with other word vectors is found to correlate significantly with humor. Table 1 shows the Pearson correlation coefficients of the predicted humor ratings with the actual EH humor ratings on the 4,997 EH words. The first column uses all EH words for fitting the linear model and for evaluation but is prone to overfitting since the models have 300 dimensions. For the second column, 10-fold cross validation is used to predict the labels (so each fold is predicted separately as hold-out) and then the correlation between the predictions and EH ratings is computed. Finally, the process is repeated 100 times to compute mean correlations and 95% confidence intervals.

Embedding name	Corr.	Corr. (using CV)
GNEWS (word2vec)	0.721	0.675 ± 0.003
WebSubword (fastText)	0.767	0.729 ± 0.002
WebFast (fastText)	0.767	0.730 ± 0.002
WebGlove (gloVe)	0.768	0.730 ± 0.002

Table 1. Correlation coefficients compared across embeddings using least-squares regression to predict EH humor ratings. Correlations with 10-fold cross-validation and confidence intervals also reported.

We find that the vectors trained on the Web crawls slightly outperform GNEWS. This may be because of improved
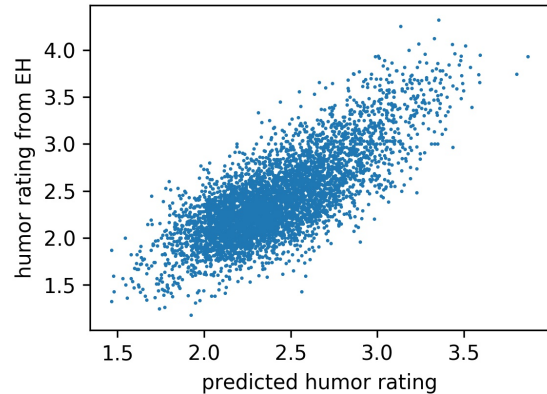

Figure 1. Humor rating predictions from a linear regression model trained on the WebSubword embedding vectors of words from the EH dataset, plotted against the EH mean humor ratings. See Table 1 for comparison with other embeddings.

training algorithms but more likely the fact that they all use at least sixfold more training data. (We also note that a number of additional fastText and gloVe embeddings, trained on different and generally smaller corpora, were found to have smaller humor rating correlations, presumably due to the fewer tokens for training.) Also consistent with prior work (Bojanowski et al., 2017), the addition of subword information does not significantly change the performance of the fastText model. However, it does conveniently make it possible to compute embeddings of all words and phrases and thus compare it to the GNEWS embedding on the larger dataset (see Section 6).

Although the set of 5k words used in this evaluation is somewhat limited, it is a verified dataset used in a previous study of single word humor. Additionally, since the words are common to all the embeddings, it allows us to draw a comparison across embeddings. As we will show later, in the broader data collection task we preformed we found that many of the words which were rated as funniest were not included in EH and are less commonly used. Nonetheless, linear directions in the word embeddings correlated with the differences in humor ratings across words.

While we see statistically significant correlations between humor vectors and EH ratings in all evaluated embeddings, several questions remain. First, do WEs capture the rich structure in the differences between individuals' senses of humor? This cannot be answered from the EH data as it only reports aggregate mean ratings, not to mention that it does not contain many of the funniest words. A second and related question is, what is the nature of this correlation? For example, it could be the case that only one topic such as words with sexual connotations are both rated as highly humorous and captured by a linear direction in the word

Table 2. Raters were clustered based on their sense-of-humor embeddings, i.e., the average vector of the words they rated funny. For each cluster, the single words that differentiate each rating cluster are shown together with scores on six humor features and demographics. Note that the demographics differences emerged from clustering the sense-of-humor vectors—demographics were not used during clustering.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
word 1	gobbledegook	tootsies	clusterfuck	whakapapa	dickheads
word 2	kerfuffle	fufu	batshit	codswallop	twat
word 3	hullaballoo	squeegee	crapola	dabbawalas	cocking
word 4	razzmatazz	doohickey	apeshit	pooja	titties
word 5	gazumped	weenies	fugly	spermatogenesis	asshattery
word 6	boondoggle	muumuu	wanker	diktats	nutted
word 7	galumphing	thingies	schmuck	annulus	dong
word 8	skedaddle	wigwams	arseholes	chokecherry	wanker
word 9	guffawing	weaner	dickheads	piggeries	cockling
word 10	bamboozle	peewee	douchebaggery	viagogo	pussyfooting
sound	**1.11**	1.02	0.97	1.02	0.90
scatological	0.80	0.99	**1.15**	0.89	1.14
colloqial	0.95	1.00	**1.14**	0.87	1.02
insults	0.86	0.90	**1.23**	0.84	1.12
juxtaposition	0.89	0.86	0.99	1.10	**1.13**
sexual	0.81	0.91	0.99	1.00	**1.25**
female %	70.3%*	57.5%	53.8%	52.4%	35.2%*
mean age	38.6	37.4	42.3*	37.2	34.7*

Bolded numbers represent the highest mean value obtained for that feature in any of the clusters.
*statistically significant difference with p-value $< 10^{-6}$.

embedding, perhaps due to the fact that they often appear in similar contexts. To what extent are more subtle features such as humorous sounding words also captured? These questions motivated the collection of our crowdsourced datasets, focusing on differences in individual-based humor ratings and humor feature annotations.

5. Clustering Sense-of-Humor Embeddings

It is well known that humor differs across groups, cultures, and individuals. We hypothesize that an individual's sense of humor can be successfully embedded as a vector as well. More specifically, if each person's "sense-of-humor" embedding is taken to be the vector average of words they rate as funny, this may predict which new, unrated words different people would find funny.

For exploratory purposes, we next clustered the raters based on their sense-of-humor embeddings, i.e., the normalized average GNEWS vector of the 36 words they rated positively. We used K-means++ in sci-kit learn with default parameters (Pedregosa et al., 2011).

For each cluster i and each word w, we define a relative mean $\mu_i(w)$ for that cluster to be the difference between the fraction of raters in that cluster who rated that word positively and the overall fraction of raters who rated that

word positively. Table 2 shows, for each cluster, the ten single words with maximum relative mean for that cluster (phrases are not displayed to save space). Table 2 also shows, for each feature of humor, the mean feature score of words rated positively by the individuals in that cluster, normalized by the overall mean. The number of clusters was chosen to be $K = 5$ using the elbow method (Thorndike, 1953), though different K values, different embeddings, and different random seeds exhibit qualitatively similar clusters.

The emergent demographic differences between the clusters that emerged are surprising, given that the clusters were formed using a small number of word ratings (not demographics). First, Cluster 1 is significantly female-skewed (compared to overall 53.4% female), Cluster 5 is male-skewed and younger, and Cluster 3 is older. How statistically significant are the age and gender differences between the clusters? We estimate this by randomly shuffling and partitioning the users into groups with the same sizes as the clusters repeatedly 10^7 times and computing statistics of the cluster means. From this, the 95% confidence interval for mean age was $(36.9, 39.2)$ and for percentage female was $(48.5, 58.2)$. Hence, the significant age differences were clusters 3 and 5 and the significant gender differences were in clusters 1 and 5. Moreover these four differences were greater than any observed in the 10^7 samples.

Table 2 also exhibits the mean feature ratings of the words rated funny by cluster members, normalized by the overall mean feature rating. Interesting differences emerged. Consistent with table 7 (see appendix), strings in the female-dominated cluster (Cluster 1) were rated as more "funny-sounding" while strings in the male-dominated cluster (Cluster 5) scored higher on sexual connotation and juxtapositions. The "older" cluster 3 rated words with scatological connotations as more humorous as well as insults and informal words. While none of the humor features we mentioned stand out for clusters 2 or 4, the clustering suggests new features that may be worth examining in a further study. For instance, Cluster 4 appears to highly rate unfamiliar words that may be surprising merely by the fact that they are words at all. Cluster 2 seems to rate "random" nouns highly (see the analysis of *concreteness* in Engelthaler & Hills (2017)) as well as words like "doohickey" and "thingie" which can be described as "random" nouns.

Table 7 (see appendix) reports the strings that males and females differed most on, sorted by confidence in the difference. A key difference between Tables 2 and 7 is whether or not the WE was used — in Table 2 the participants were clustered by their sense of humor embeddings without using their gender, while Table 7 presents differences in word ratings without using the WE at all. It is interesting that the WE recovers gender with a nontrivial degree of accuracy from the sense-of-humor embeddings. In the next section, we discuss predicting humor features, many of which emerged naturally in the clusters, using word embedding vectors.

6. Humor Features in the Embedding

Recall that workers rated the six theoretically-grounded features of humor described in Section 2 on 1,500 words (including the funniest 216 word set), as described in section 3.3. Using supervised learning, we compute each feature's "predictability" as follows. As in the previous section, we use least squares linear regression to fit a linear function to the feature values as a function of the 300-dimensional embeddings, and use 10-fold cross validation to form predictions of feature tagging on all words. As before, we then compute the correlation coefficient between the predictions and labels. This is repeated 100 times and the mean of the correlation coefficients is taken as the feature predictability. To compute feature correlation with humor, we use the mean ratings of the 216 words in the personalized data and assign a rating of -1 to any word not in that set.

Figure 2 shows these correlation with mean humor rating (y-axis) versus predictability (x-axis) across the words embeddings GNEWS and WebSubword. The GNEWS representation is well-suited for predicting colloquial (informal) words and insults, with correlations greater than 0.5, while the feature that was most difficult to predict with GNEWS
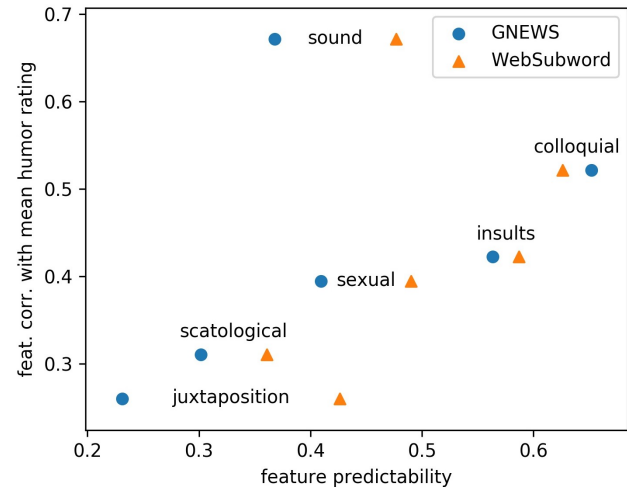

Figure 2. Feature correlation with mean humor ratings vs. feature predictability (word-embedding correlations, based on GNEWS and WebSubword) for the six features on the 1,500 words with feature annotations.

was the juxtaposition feature, with a correlation slightly greater than 0.2. All the features had positive correlation with mean humor ratings, with funny sounding having the highest correlation.

WebSubword predicted all features better than GNEWS with the exception of the colloquial feature. WebSubword yielded correlations which were equal or greater than 0.4 for five of the six identified humor features. The greatest difference was seen in the predictability of the juxtaposition feature, perhaps due to WebSubword's ability to extract meaningful information from subwords, highly relevant to a juxtaposed word which combines two shorter terms.

We can easily evaluate predictability and correlations with humor for an automatic feature such as word length. For word length, we find a predictability correlation coefficient of 0.518 indicating a good ability of a linear direction in the WE to predict word length, and a correlation with mean humor ratings of -.126, consistent with EH's findings that shorter words tend to be rated higher.

7. Predicting Individual Differences in Word Humor

We can use supervised learning to evaluate our earlier hypothesis that the average of the embeddings of the words a person finds funny captures their sense of humor in vector form. We consider the following test inspired by the comedy maxim, "know your audience." In this test, we take two random people p_1 and p_2 with funniest words w_1, w_2 (rating of 3), respectively, where we require that p_1 rated w_2 as 0 and p_2 rated w_1 as 0. Note that this requirement is satisfied

for 60% of pairs of people in our data. In other words, 60% of pairs of people had the property that neither one chose the other's funniest word even in the first round (of the 216 words study). This highlights the fact that individual notions of humor indeed differ significantly.

Given the two sets of 35 other words each participant rated positively, which we call the *training words*, and given words $\{w_1, w_2\}$ which we call the *test words*, the goal is to predict which participants rated w_1 or w_2 funniest. Just as a good comedian would choose which joke to tell to which audience, we use the WE to predict which person rated which test word as funny, based solely on the training words. For example, the training sets might be $\{$*bamboozle, bejeezus, …, wigwams, wingnuts*$\}$ and $\{$*batshit, boobies, …, weaner, whakapapa*$\}$ and test set to match might be the words $\{$*poppycock, lollygag*$\}$.

Table 3. Success rates at know-your-audience tests, which test the ability of sense-of-humor embeddings to distinguish different raters based on sense of humor and predict which of two words each would find funny.

Know-your-audience test	Success rate
Easy: disjoint sets of 35 training words.	78.1%
Normal: 35 training words.	68.2%
Hard: 5 training words.	65.0%

To test the embedding we simply average training word vectors and see which matches best to the test words. In particular, if $\vec{r}_1$ and $\vec{r}_2$ are the two training word vector averages and $\vec{v}_1$ and $\vec{v}_2$ are the corresponding test word vectors, we match $\vec{r}_1$ to $\vec{w}_1$ if and only if,

$$\|\vec{r}_1 - \vec{v}_1\|^2 + \|\vec{r}_2 - \vec{v}_2\|^2 < \|\vec{r}_1 - \vec{v}_2\|^2 + \|\vec{r}_2 - \vec{v}_1\|^2$$
$$0 < (\vec{r}_1 - \vec{r}_2) \cdot (\vec{v}_1 - \vec{v}_2)$$

Simple algebra shows the above two inequalities are equivalent. Thus, the success rate on the test is the fraction of eligible pairs for which $(\vec{r}_1 - \vec{r}_2) \cdot (\vec{v}_1 - \vec{v}_2) > 0$. Note that this test is quite challenging as it involves prediction on two completely unrated words, the analog of what is referred to as a "cold start" in collaborative filtering (i.e., predicting ratings on a new unrated movie).

We also consider an easy and a hard version of this know-your-audience test. In the easy version, a pair of people is chosen who have disjoint training sets of positively rated words, indicating distinct senses of humors. In the hard version, we use as training words only the five words each person rated 2. It is important to note that no ratings of the test words (or the other participants) are used, and hence it tests the WE's ability to generalize (as opposed to collaborative filtering). The test results given in Table 3 were calculated by Brute force over all eligible pairs in the data

(1,004 for the easy test and 818,790 for the normal and hard tests).

8. Conclusions

Can NLP-humor be closer than we think? This work suggests that many aspects of humor, including individual differences, may in fact be easier than expected to capture using machine learning – in particular many correlate with linear directions in Word Embeddings. In particular, each individual's sense-of-humor can be easily embedded using a handful of ratings, and that differences in these embeddings generalizes to predict different ratings on unrated words. We have shown that word humor possesses many features motivated by theories of humor more broadly, and that these features are represented WEs to varying degrees.

WEs are of course only one lens through which single word humor can be studied. Phonotactic, morphological and orthographic characteristics of humor can certainly be used to complement the WE approach. However, in this study, we show that WEs, popular NLP tools, are already enough to achieve meaningful progress in explaining single word humor. Furthermore, we offer our originally collected datasets, which we have made publicly available, as a useful contribution to future projects looking to complement and extend ours.

There are numerous possible applications of word humor to natural language humor more generally. As discussed, comedians and writers are aware and indeed use such word choice to amplify humorous effects in their work. Our findings could therefore contribute to building aiding tools for creative writing tasks. Similarly, humorous word detection may help in identifying and generating humorous text. Moreover, our ratings could be used by text synthesis systems such as chat-bots which use WEs to tweak them towards or away from different types of humor (e.g., with or without sexual connotations), depending on the application at hand and training data.

Finally, one approach to improving AI recognition and generation of humor is to start with humorous words, then move on to humorous phrases and sentences, and finally to humor in broader contexts. Our work may be viewed as a first step in this programme.

References

Barbieri, F. and Saggion, H. Modelling irony in twitter. In *EACL*, pp. 56–64, 2014.

Beard, R. *The 100 Funniest Words in English*. Lexiteria, 2009. ISBN 9780615267043.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017.

Bown, L. *Writing Comedy: How to use funny plots and characters, wordplay and humour in your creative writing*. Hodder & Stoughton, 2015. ISBN 9781473602205. URL https://books.google.com/books?id=HPzMBAAAQBAJ.

Chandrasekaran, A., Vijayakumar, A. K., Antol, S., Bansal, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. We are humor beings: Understanding and predicting visual humor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4603–4612, 2016.

Chen, P.-Y. and Soo, V.-W. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, pp. 113–117, 2018.

Cooper, A. A. Sensus communis: An essay on the freedom of wit and humour. *Characteristicks of men, manners, opinions, times*, 1, 1709.

Engelthaler, T. and Hills, T. T. Humor norms for 4,997 english words. *Behavior Research Methods*, Jul 2017. ISSN 1554-3528. doi: 10.3758/s13428-017-0930-6. URL https://doi.org/10.3758/s13428-017-0930-6.

Freud, S. Jokes and their relation to the unconscious (j. strachey, trans.). *N. Y.: Holt, Rinehart, and Winston*, 1905.

Hossain, N., Krumm, J., Vanderwende, L., Horvitz, E., and Kautz, H. Filling the blanks (hint: plural noun) for mad libs humor. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 638–647, 2017.

Joshi, A., Tripathi, V., Patel, K., Bhattacharyya, P., and Carman, M. Are word embedding-based features useful for sarcasm detection? *arXiv preprint arXiv:1610.00883*, 2016.

Kant, I. *Critique of judgment*. Hackett Publishing, 1790.

Kierkegaard, S. Concluding unscientific postscript to philosophical fragments (translation by d. swenson, 1941), 1846.

Kortum, R. D. Funny words: The language of humor. In *Varieties of Tone*, pp. 134–142. Springer, 2013.

Mihalcea, R. and Strapparava, C. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 531–538, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1220575.1220642. URL https://doi.org/10.3115/1220575.1220642.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, 2013b.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Morreall, J. Philosophy of humor. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

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

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Radev, D., Stent, A., Tetreault, J., Pappu, A., Iliakopoulou, A., Chanfreau, A., de Juan, P., Vallmitjana, J., Jaimes, A., Jha, R., et al. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. *arXiv preprint arXiv:1506.08126*, 2015.

Raz, Y. Automatic humor classification on twitter. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pp. 66–70. Association for Computational Linguistics, 2012.

Schopenhauer, A. *Arthur Schopenhauer: The World as Will and Presentation: Volume I (translation by R. B. Haldane, 1888.* Routledge, 1844.

Shahaf, D., Horvitz, E., and Mankoff, R. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1065–1074. ACM, 2015.

Simon, N. *The Sunshine Boys: A Comedy in Two Acts*. Samuel French, Inc., 1974.

Spencer, H. *Essays on Education, Etc.* London: Dent, 1911.

Stock, O. and Strapparava, C. Hahacronym: Humorous agents for humorous acronyms. *Humor*, 16(3):297–314, 2003.

Strauss, U., Grzybek, P., and Altmann, G. Word length and word frequency. In *Contributions to the science of text and language*, pp. 277–294. Springer, 2007.

Thorndike, R. L. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

tvtropes.org. *Just For Fun / Inherently Funny Words*, 2017. http://tvtropes.org/pmwiki/pmwiki.php/JustForFun/InherentlyFunnyWords, accessed 2017-07-25.

Van Hee, C., Lefever, E., and Hoste, V. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 39–50, 2018.

West, R. and Horvitz, E. Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances". In *Proceedings of AAAI 2019*, 2019.

Westbury, C., Shaoul, C., Moroschan, G., and Ramscar, M. Telling the worlds least funny jokes: On the quantification of humor as entropy. *Journal of Memory and Language*, 86:141 – 156, 2016. ISSN 0749-596X. doi: https://doi.org/10.1016/j.jml.2015.09.001. URL http://www.sciencedirect.com/science/article/pii/S0749596X15001023.