# Inside Jokes: Identifying Humorous Cartoon Captions

Dafna Shahaf
Microsoft Research
dshahaf@microsoft.com

Eric Horvitz
Microsoft Research
horvitz@microsoft.com

Robert Mankoff
The New Yorker Magazine
bob_mankoff@newyorker.com

## ABSTRACT

Humor is an integral aspect of the human experience. Motivated by the prospect of creating computational models of humor, we study the influence of the language of cartoon captions on the perceived humorousness of the cartoons. Our studies are based on a large corpus of crowdsourced cartoon captions that were submitted to a contest hosted by the New Yorker. Having access to thousands of captions submitted for the same image allows us to analyze the breadth of responses of people to the same visual stimulus.

We first describe how we acquire judgments about the humorousness of different captions. Then, we detail the construction of a corpus where captions deemed funnier are paired with less-funny captions for the same cartoon. We analyze the caption pairs and find significant differences between the funnier and less-funny captions. Next, we build a classifier to identify funnier captions automatically. Given two captions and a cartoon, our classifier picks the funnier one 69% of the time for captions hinging on the same joke, and 64% of the time for any pair of captions. Finally, we use the classifier to find the best captions and study how its predictions could be used to significantly reduce the load on the cartoon contest's judges.

## Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors; H.2.8 [**Database Applications**]: Data mining; H.3.1 [**Content Analysis and Indexing**]: Linguistic processing

## Keywords

Humor, Cartoon, Cartoon Caption

## 1. INTRODUCTION

Humor plays a central role in the lives and minds of people. A considerable amount of discussion has been devoted to the nature and function of humor. Multiple hypotheses have been proposed, from models about the deep roots of humor in human cognition to its role in socialization. To date, most investigations of humor have been undertaken within psychology, philosophy, and linguistics; in comparison, computational research on humor is still in its infancy.

Creative pieces, such as jokes, are traditionally considered to be outside the reach of computer science. Humor is viewed as a distinctly human trait; even in science fiction, robots and computers are almost always portrayed as humorless – no matter how proficient they are with language or other skills.

We believe that understanding humor defines an intriguing set of challenges for computer science. Endowing machines with capabilities for recognizing and generating humor could enhance human-computer interaction and collaboration, through improved understandings of semantics, intentions, and sentiment. Recently, inability to detect humor led to a failure in algorithmic-trading systems: shares of Tesla Motors rose and trading volume surged after a clearly parodic April Fool's press release [1].

| |
|---|
| What's it going to take to get you in this car today? |
| Relax! It just smells the other car on you. |
| It runs entirely on legs. |
| Just don't tailgate during mating season. |
| It's only been driven once. |
| He even cleans up his road kill. |
| The spare leg is in the trunk. |
| Comfortably eats six. |
| She runs like a dream I once had. |

**Figure 1: Example cartoon from the New Yorker contest, with the shortlist of submitted captions.**


*KDD'15,* August 10-13, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2783388.

Humor can also be harnessed to increase attention, retention, and engagement, and thus has numerous interesting applications in education, health, communications, and advertising. Beyond applications, pursuing information-theoretic models of humor could lead to new insights about the use of language and about deeper foundations of human cognition.

Computational work to date on humor focuses largely on humor generation in limited domains, such as puns and humorous acronyms [5, 29]. Several other projects focus on the related task of humor recognition [23, 30, 32].

In this paper, we describe a new direction in the realm of humor recognition. We focus on the task of identifying humorous captions for cartoons. Specifically, we consider cartoons from The New Yorker magazine. The New Yorker holds a weekly contest in which it publishes a cartoon in need of a caption. Readers are invited to submit their suggestions for captions. The judge selects a shortlist of the funniest captions, and members of the editorial staff narrow it down to three finalists. All three finalists' captions are then published in a later issue, and readers vote for their favorites. Figure 1 shows an example cartoon. In this cartoon, a car salesperson attempts to sell a strange, hybrid creature that appears to be part car, part animal. The judge's shortlist appears under the cartoon in Figure 1.

The absence of large humor corpora has perhaps held data-driven humor research back. The New Yorker cartoon contest corpus is especially intriguing in this regard, as it contains thousands of captions for the same image. This allows us to analyze responses of people to the same visual stimulus, therefore enabling testing of hypotheses that would be hard to test on other corpora (e.g., joke collections).

**Our contributions are as follows:** We aim to predict the relative humorousness of captions without deep image analysis and text understanding, by leveraging human tagging of scenes and automated analysis of linguistic features.

To test our ideas, we create a dataset of crowdsourced captions from the New Yorker competition, along with human judgments. We identify and analyze different variations of the same joke and find factors affecting the level of perceived humor. We also automatically identify the joke within a caption, and quantify the intensity of humor of different jokes and their relation to the cartoon.

We formulate a pairwise evaluation task and construct a classifier that, given two captions and a cartoon, determines which caption is funnier. Our classifier achieves 69% accuracy for captions hinging on the same joke, and 64% accuracy comparing any two captions. We implement a Swiss-system tournament that ranks all captions. On average, all of the judges' top-10 captions are ranked in the top 55.8%, thereby suggesting that the methods can be used to significantly reduce the workload faced by judges.

Beyond the details of the current study, we seek more broadly to frame data-centric research on humor. We wish to share directions for joining data mining with prior work on the psychology of humor for exploring an important dimension of digital content. We believe that data mining should and will be applied to psychological phenomena and that there are numerous unexplored opportunities at the intersection of these fields.

# 2. EFFECT OF CAPTION PHRASING

We obtained 16 New Yorker cartoons, along with all of their submitted captions. Each cartoon is associated with a set of captions ranging in size from 3562 to 6557 captions, averaging 4808 captions per cartoon. Based on this data, we formulate a pairwise comparison task: Given a cartoon and two captions, determine **which is funnier.**

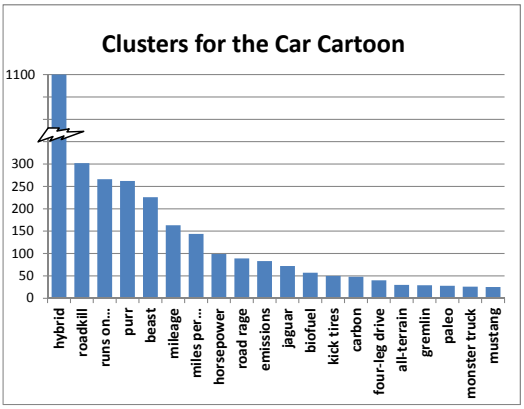



**Figure 2: Cluster sizes for similar jokes for the cartoon in Figure 1. 1096 people submitted a joke hinging on the car being a "hybrid".**

There are many reasons that people could perceive one caption as funnier than another. Most importantly, different captions may tell completely different jokes. To explore the *linguistic* properties that make a caption funny, we first focused our attention on captions that make the same fundamental joke.

Although the participants work independently, they tend to produce surprisingly similar ideas in response to a cartoon stimulus. For example, consider again the car cartoon in Figure 1. Figure 2 shows the number of people submitting similar ideas for this cartoon. 1096 people – 17% of all the participants – submitted a joke hinging on the car being a "hybrid". Over three hundred people submitted some variant of "it runs on roadkill". Other popular topics included "don't kick the tires", "purrs like a kitten", "the spare leg is in the trunk", "carbon footprint"/"biofuel", "horsepower" (see Table 1 for a small sample), "all-terrain vehicle" and "she's really a beast" – each one repeating tens of times. This situation is common across the cartoons.

## 2.1 Acquiring Judgments

We picked 10 cartoons from the dataset. Our first goal was to extract clusters of captions that leverage the same

| |
|---|
| We don't even talk about horsepower. |
| Horsepower? Who talks about horsepower anymore? |
| With this baby, who needs horsepower? |
| We don't measure it in horsepower. |
| Think of it like horsepower, redefined. |
| Four for breakfast, that's where the horsepower comes in |
| And you thought horsepower was the way to go. |
| Oh, horsepower is such a old term with cars.... |
| Actually, the term horsepower is meaningless with this baby. |
| It has more horse power than you can imagine! |

**Table 1: Sample batch of captions coming from a single cluster**

joke concept. King et al. have tackled the problem of clustering short discourse units on a similar cartoon-caption dataset [20]. Their results suggest a simple clustering algorithm that is very effective at finding coherent factoids. We note that they were interested in clustering all (or most) of the captions, while we focus on an easier task: we are only concerned with the precision of clustering jokes into a few large groups. We clustered the captions and picked the three largest clusters for each cartoon.

Next, we randomly selected 10 entries from each cluster. Every ten entries constitute a single batch (see Table 1 for a sample batch). We controlled for the length of captions given a sense that length could have a significant effect on perceived funniness of captions. We hoped that controlling for length would bring into focus other, potentially more subtle factors.

To control for length, we ensure that each batch includes five captions of length 5-6 words and five of length 9-10 words. Later, we will only perform comparisons between captions of similar length; two different groups of lengths were chosen in order to capture different levels of complexity. The specific lengths (5-6, 9-10) were chosen since they represent many of the shortlisted captions.

To understand what makes some captions funnier than others, we first needed to obtain humor judgments. We recruited crowdworkers via Mechanical Turk and challenged the workers with a task of comparing the funniness of the captions. The workers were presented with a cartoon and five captions, drawn at random from the same batch. They proceeded to sort the captions from the funniest to the least funny. Each answer provided us with 10 pairwise comparisons (#1 is funnier than #2, #3, #4 and #5. #2 is funnier than #3, #4 and #5 and so on).

Other studies of humor have asked participants to rate jokes on a Likert scale. Humor judgments have been obtained by averaging the assessed Likert scores. We note that there is a long-running dispute about whether it is valid to average Likert scores. For our study, we do not believe we can treat ordinal rating data as interval data. Psychological studies on subjective evaluations [31] suggest that the pairwise setup is preferred to direct rating, and thus we chose to ask the workers for direct comparisons.

We selected pairs of captions that achieved high agreement among the rankers (80% agreement or more, similar length, ranked by at least five people). These pairs served as our ground truth. We had 30 batches, each one ranked by 30-35 workers, for a total of 1016 tasks and 10,160 pairwise rankings. Only 35% of the unique pairs that were ranked by at least five people achieved 80% agreement, resulting in 754 pairs. This indicates the hardness of the task.

## 2.2 Hypotheses: Indications of Humor

To formulate the task as a machine learning problem, we sought features of captions that could help to distinguish between funny and less funny captions. Later, we evaluate and compare the predictive power of these features.

### 2.2.1 Unusual Language

For one set of features, we pursued the idea that captions deemed funnier might tend to employ language in a more unusual manner. In order to quantify "unusual", we take a language-model approach. A language model is a function that specifies a probability measure over strings. We train language models on 2GB of ClueWeb data [13], meant to represent common language. To assess different levels of distinctiveness, we compute four models, using 1-grams, 2-grams, 3-grams, and 4-grams respectively.

After computing the language models, we measure the *perplexity* of each caption in the context of each language model. The perplexity is the exponential of the negative average log-likelihood. In other words, the perplexity of a caption is an inverse function of how well our language model predicts it: the higher the perplexity, the more distinctive or "stand out" the caption is.

| **Feature** | **%Funnier is higher** |
|---|---|
| Perplexity (1-gram) | 0.45** |
| Perplexity (2-gram) | 0.49 |
| Perplexity (3-gram) | 0.46* |
| Perplexity (4-gram) | 0.46* |
| POS Perplexity (1-gram) | 0.45* |
| POS Perplexity (2-gram) | 0.52 |
| POS Perplexity (3-gram) | 0.5 |
| POS Perplexity (4-gram) | 0.51 |
| Sentiment | 0.61* |
| Readability | 0.57**, 0.56* |
| Proper Nouns | 0.48 |
| Indefinite articles | 0.43 |
| 3rd Person | 0.58 |
| Location (quarters) | 0.53, 0.42*, 0.49, 0.55* |

**Table 2: Same joke: Percentage of caption pairs in which a feature has higher numerical value in the funnier caption (higher readability score, more proper nouns, etc). Significance according to a two-sided Wilcoxon signed rank test is indicated using *-notation (*p≤0.05, **p≤0.005, Holm-Bonferroni correction)**

Table 2 shows the results of the language model analysis. Interestingly, funnier captions tend to use less-distinctive language (recall that we are only comparing captions that tell the same joke). We note that "keep it simple" seems to be one of the most common pieces of advice given to comedy writers, and this recommendation appears explicitly in other cartoon caption contest rules.

In addition to computing lexical distinctiveness, we compute the *syntactic* distinctiveness of funny captions. To do this, we replace each word in the captions and our web corpus by its corresponding part-of-speech (POS) tag, obtained using NLTK[1]. We then learn language models and compute perplexity scores on the POS corpus.

Funny captions tend to have less distinct grammar on a unigram level; the results for higher $n$-gram models were not significant (see Table 2). This again suggests that simpler captions are preferred by the crowdworkers who had judged the funniness of the captions.

### 2.2.2 Sentiment

In our search for useful features for discriminating funnier captions, we found research on humor by Mihalcea and Pulman [23]. In this work, the authors explore the distinctive linguistic features of humor. One of the characteristics they

[1]http://www.nltk.org/

found was *negative orientation*: a frequent use of negative words (cannot, bad, illegal, mistake) in humorous texts.

In this spirit, we used an in-house sentiment analysis tool to annotate the captions. Surprisingly, our data suggests the opposite to the findings of Mihalcea and Pulman, as funnier captions tended to be more positive (see Table 2). This finding may be due to the crowdworkers humor preferences, the nature of the content at hand, or to the sentiment analyzer's inability to detect the subtleties of sarcasm associated with captions for the cartoons.

#### 2.2.3 Taking Expert Advice

Patrick House, a winner of the New Yorker contest, wrote an article about his winning strategy [15]. He suggests to use "common, simple, monosyllabic words" and to "Steer clear of proper nouns that could potentially alienate".

Following this advice, we measure readability using Flesch reading ease [12] and Automated Readability index [27]. These were designed with longer texts in mind, but we believe they could be useful for us. We also added a binary feature for the presence of any proper nouns in the caption.

In a similar vein, recent work [10] postulated that movie quotes are more memorable when they can be applied in multiple different contexts. To obtain a proxy for the generality of a caption, we measured the number of third-person pronouns and the number of indefinite vs. definite articles. Third-person pronouns refer to something that was introduced earlier, and are therefore harder to use in new contexts. Similarly, indefinite articles are more likely to refer to general concepts than definite articles.

Table 2 shows the results. Funnier captions achieved better readability scores, but the rest of the features did not show discriminatory power. As we were comparing instances of the same conceptual joke, several of these features did not vary significantly between the alternative captions. For example, consider proper nouns: the occasional biofuel-inspired joke might mention a Prius (or give the characters a first name), but this was a rare event.

#### 2.2.4 Location of Joke Phrase

Finally, we wanted to know whether the location of the joke within the caption makes a difference. For example, does it help to have a humorous ending, providing a surprising or witty punch line? Since we know the cluster that each caption came from, we can easily identify the location of the joke in each caption.

We added four features, indicating which quarter of the caption contains the joke phrase. The results indicate that less-funny captions tend to have the joke in the second quarter more often, and the funnier captions tend to defer the revelation of the core joke to the end (Table 2).

### 2.3 Summary of Results

We find strong evidence that, from a lexical perspective, funnier captions are simpler than their less funny counterparts – lexically, syntactically, and using ease-of-reading metrics. There is some evidence that less funny captions present the core joke earlier in the caption than funnier ones.

Other metrics, such as proper nouns and features capturing generality, did not achieve significant differences. This might be due to comparing paraphrases of the same joke, or perhaps the size of the data set; we will return to these features in the next sections.

### 2.4 Prediction Task

We now show how the analyses and attributes discussed above can provide features for a basic prediction task, similar to the one that we presented our human judges with: given a pair of captions, identify the caption that is funnier.

Given a pair of captions, we extract the features discussed in Section 2.2 from both. We construct a vector from both sets of features, along with the *differences* between them. The label of the vector indicates whether the first caption is funnier than the second.

We perform a 5-fold cross-validation, as the size of the dataset can make 10-fold too sensitive. The split is done based on the cartoon, so the algorithm is always tested on cartoons it was not trained on. This is meant to simulate real usage by the New Yorker editor, judging captions for a new contest. Since we can make each pair of captions into a positive or a negative example (by changing the ordering), it was easy to balance the dataset, and thus the random baseline accuracy is 50%.

We first try to predict the funnier caption using a standard bag-of-words model. That is, we try to predict which caption is funnier based on content alone. A random forest using only bag-of-words features achieves 52% accuracy, only slightly outperforming the random baseline.

We evaluate the importance of different features using the Gini importance score [6]. The Gini importance of a feature is the normalized total reduction of the criterion brought by that feature. Top features, identified by averaging across the decision trees, can be grouped into the following classes:

**Pronouns:** you, I, it, my, we, me, your, their

**Question words:** how, what

**Negation words:** *n't, no, not

**Auxiliary verbs:** do, should, need, can, (think)

The result that these words were top discriminators is surprising, given that many of these words are usually considered discardable "stopwords" (non-informative words which should be filtered out in pre-processing) in natural language analyses. Interestingly, another characteristic of humorous texts identified in [23] (in addition to negative polarization) is human-centric vocabulary, including pronouns. Also, we note that no explicit joke words achieved high importance as discriminators, most likely because we compare captions that tell the same joke. The first proper noun to appear in that list is 'Google' (32nd place), which is not a joke word in any of the clusters.

After evaluating the bag-of-words classifier, we added the features discussed in Section 2.2. With all features included, a random forest classifier achieves 69% accuracy. The most important features included perplexity (both lexical and POS), sentiment and readability. No bag-of-word feature was ranked in the top 50 important features.

## 3. COMPARING DIFFERENT JOKES

After exploring different variations of the same joke, we now move on to comparing *different* jokes. Comparing different jokes is a much harder task, and potentially requires understanding deeper, more universal underpinnings of humor.

**Context:** office, workplace, secretary, phone
**Anomaly:** stairway, sky, heaven

**Figure 3: Cartoon from the New Yorker contest and the tags associated with it. One list describes the general setting, or context, of the cartoon. The other describes its anomalies.**

There are three main theories of humor [3]: relief theory (humor releases nervous energy or tension), superiority theory (humor allows one to feel superior), and incongruity theory (humor is about violating our mental patterns and expectations). A more recent theory suggests that humor stems from violations that are simultaneously seen as benign, or harmless [22].

Incongruity theory is currently the dominant theory of humor in philosophy and psychology [25]. The idea that incongruity is central to humor is centuries old; however, there is no rigorous definition that would allow a researcher to detect incongruity in a given situation.

In the case of cartoons, the image itself is incongruous, going against our expectations. This incongruity is often the result of unusual combinations. Hofstadter et al. referred to a similar phenomenon as "frame blends" [14]. A frame blend occurs when a person blurs two distinct situations or scenarios in their mind, creating a hybrid situation. Virtually all of the New Yorker cartoons fall under this category, e.g., a car with animal legs (Figure 1), a baseball player wearing high heels, a man sitting inside a cage at a dog kennel, and so on. In these cartoons, the caption often attempts to resolve the surprising incongruity. For a cartoon to be funny, the reader needs to understand both the picture and the caption, and then to integrate them into a meaningful whole. When the pieces fall into place, readers feel satisfaction and the relieved tension is expressed as humor.

The incongruity theory for cartoon humor is further supported by psychology research: Jones et al. analyzed the components of cartoon humor and determined the effects of caption, picture, and the interaction between them on the humor rating of the whole cartoon [17]. They found that the humor of the caption was not correlated significantly with the humor rating of the entire cartoon. However, the *interaction* between picture and caption was highly correlated with the humor rating of the cartoon. Interestingly, the funnier the picture and the *less funny* the caption, the highest the cartoon was ranked. This result may be related to the findings that we presented in the previous section.

We believe that to judge a caption our algorithm must have some knowledge of the accompanying cartoon. In particular, we wish to understand the blend of frames which are in play. A growing body of research has focused on automatically generating text descriptions for images [19, 11]. However, to the best of our knowledge, nobody has successfully trained models that can extract deep semantic descriptions for cartoon data. Thus, we rely on human annotation.

For each cartoon, we asked people to provide two lists of tags, corresponding to the two blended frames that create the incongruity. One list describes the general setting, or context, of the cartoon (car, dealership, showroom, salesman). The other describes its anomalies (tires, legs, animal, combination). See Figure 3 for an example.

Since the number of cartoons is negligible compared to the number of captions, this preprocessing step is quick and easy. We obtained tags from three different annotators. We observed high inter-agreement among the annotators (average Jaccard 0.86); that is, the output was not very annotator-specific. Since we use word embeddings (see below), the exact choice of words does not matter as much.

## 3.1 Acquiring Judgments

We returned to Mechanical Turk workers for humor judgments. We composed new batches of 10 captions each, this time without requiring them to revolve around the same joke. As before, we keep pairs of captions of similar length that had at least five rankings and at least 80% agreement. We had 297 batches, each one completed by 30-35 workers, for a total of 10,007 tasks and 100,070 pairwise rankings. 45.11% of the caption pairs ranked by at least five people reached wide agreement, resulting in 4184 unique pairs.

## 3.2 Features

The features we used for the second task were nearly identical to those used for the earlier task (Section 2.2). We now explain the main differences.

### *3.2.1 Joke Location*

In Section 2.2, knowing the location of the joke in the caption was easy; the captions, after all, were picked from a cluster of captions revolving around the same idea, so the joke phrase was known. However, we no longer know the joke in advance.

We selected the word with the lowest perplexity (highest distinctiveness) according to the 4-gram model. If the word was a part of a phrase, we picked the whole phrase. While not perfect, this method was surprisingly accurate at finding the central joke in a caption. See Figure 4 for an example. The red phrases are the ones identified through perplexity, and the purple ones were out-of-vocabulary words – another useful way to identify the joke. Typos are often picked up as out-of-vocabulary words. If they happen to spell a real word, they are sometimes picked up as low-perplexity words (see, for example, the "anti-theft devise", including the misspelling of "device"). Also, note that our algorithm misses several expressions, e.g., "tiger in the tank".

Several of the captions in Figure 4 do not have a joke associated with them. These are sentences that did not have any highly distinctive words in them. For example, "And the gas mileage is incredible" is a common expression, that derives its humor only from the contrast with the cartoon. Many of the submissions (and more importantly, many of the short-

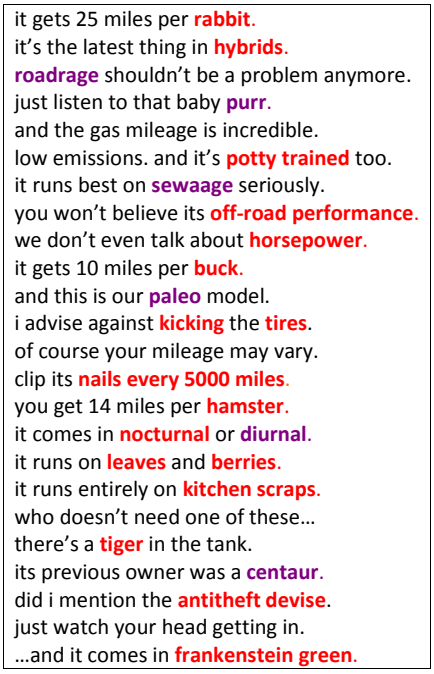



**Figure 4: Identifying joke phrases: Phrases in red have low perplexity. Phrases in purple do not appear in our vocabulary, and are often useful for finding the joke (as well as typos).**

listed captions) rely on a text that makes the extraordinary sound mundane.

Thus, in addition to the location of the joke phrase, we have added two more features: minimum perplexity across the caption, and total number of words in the joke phrase.

### 3.2.2 *Joke Geometry*

Now that we have a proxy for finding the joke phrase in a caption, we wish to measure the similarity of this phrase to the cartoon's context and the cartoon's anomalies.

Since the tags and the caption are very sparse, we needed to measure *semantic* similarity. Word embeddings have been shown to be successful at similar tasks involving capturing semantic similarities. In particular, Mikolov et al. recently introduced the Skip-gram model [24], an efficient method for learning high-quality vector representations of words from large amounts of unstructured text data. The word embeddings are computed using neural networks; the learned vectors are interesting because they seem to encode many linguistic regularities and patterns as linear translations. For example, the result of a vector calculation

$$vector(Madrid) - vector(Spain) + vector(France)$$

is closer to *vector(Paris)* than to any other word vector.

We trained word vectors on the One Billion Word Language Modeling Benchmark [8]. Using these vectors, we computed cosine similarities between the caption's main joke phrase and the cartoon tags.

Let $C, A$ be sets of context and anomaly tags, respectively, and let $J$ be the set of joke words in a specific caption (may be empty). For each word $j \in J$ we define its similarities to the cartoon. The similarity of $j$ to each set of tags is the maximum similarity (shortest distance) to any of the set's members:

$$sim(j, C) = \max_{c \in C} sim(j, c)$$

$$sim(j, A) = \max_{a \in A} sim(j, a)$$

We used four similarity scores as features:

**Context similarity:** $\max_{j \in J} sim(j, C)$

**Anomaly similarity:** $\max_{j \in J} sim(j, A)$

**Max absolute difference:**

$$\max_{j \in J} |sim(j, A) - sim(j, C)|$$

**Average absolute difference:**

$$\frac{1}{|J|} \sum_{j \in J} |sim(j, A) - sim(j, C)|$$

Figure 5 illustrates the idea. We take multiple captions submitted for the cartoon in Figure 3. We then compute their similarity to the cartoon context (office, workplace, secretary, phone) and anomalies (stairway, sky, heaven). On the left, we show the results as a scatter plot. The dashed line is $x = y$; captions above the line are more similar to the anomalies, and captions below it are more similar to the general context.

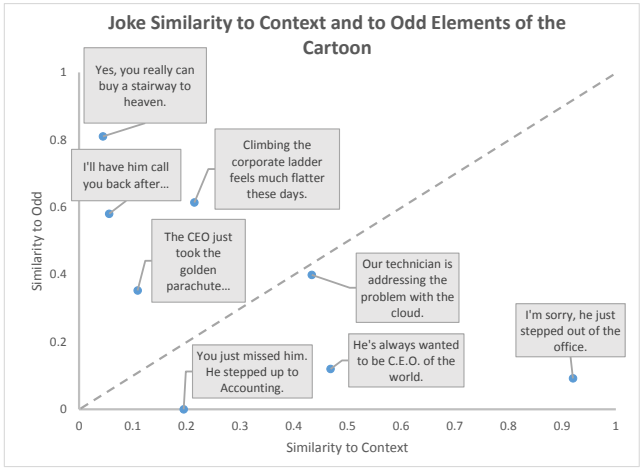
Figure 5 (right) shows the same data, but instead of displaying both similarities, it only shows the *difference* between them. This representation makes it easier to observe patterns of interest. In this representation, captions closer to the left are closer to the cartoon context. For example, "I'm sorry, he just stepped out of the office" is a perfectly reasonable sentence in an office setting. Captions on the right end are closer to the anomalies ("Yes, you really can buy a stairway to heaven"). Interestingly, wordplays and puns appear closer to the middle ("Our technician is addressing the problem with the cloud", "The CEO just took the golden parachute plus package!", "Climbing the corporate ladder feels much flatter these days").

## 3.3 Summary of Results

Table 3 shows the statistics for each of our features. It is interesting to compare these results with Section 2.3. While funnier captions still have simpler grammatical structure, it now seems like they use more distinctive vocabulary. Several features that were not significant when comparing the same joke became very significant for comparing different jokes, and vice versa.

For example, readability and sentiment are no longer as significant as they were for the single-joke task. On the other hand, proper nouns and 3rd person words have become significant, with funnier captions using significantly less of both. As mentioned earlier, same-joke captions often use similar proper nouns. The situation is different when considering multiple jokes. Many captions rely entirely on proper nouns for the joke. For example, for the cartoon in Figure 3, two of the captions read "Oh my God, Florenz Ziegfeld has gone to heaven!" and "I'm afraid Mr. Burns is gone for the afterlife." One possible explanation for the results is that names that are not universally recognizable alienate the readers.

The joke-geometry features turned out to be interesting, with absolute-differences and distance-to-context features



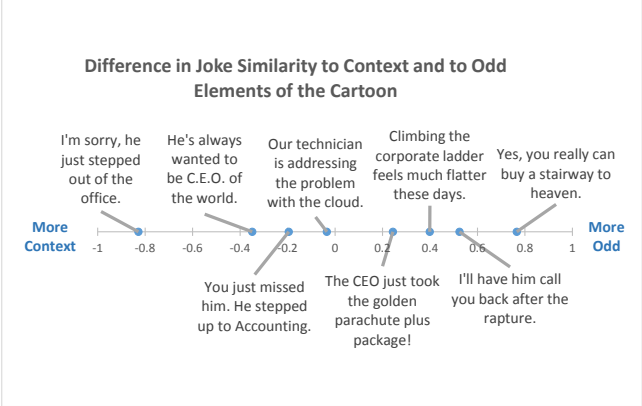




**Figure 5: We take captions submitted for the cartoon in Figure 3 and compute their similarity to the cartoon context and anomalies. On the left is a scatter plot, where each caption is a data point. The dashed line is $x = y$; captions above the line are more similar to the anomalies, and captions below it are more similar to the general context. On the right we show the same data, focusing on the *difference* between the two similarities. Captions closer to the left are closer to the cartoon context and captions on the right end are closer to the anomalies. Interestingly, wordplays and puns appear closer to the middle.**

being significant. This indicates that funny captions have more words which can "bridge" between the core context and anomalous frames in the cartoon.

Perhaps most counterintuitive is the "Length of Joke Phrase" feature. According to our data, funnier captions actually use *fewer* joke words. Many of the bland captions, where humor is achieved by making the extraordinary sound mundane ("gas mileage is incredible"), were ranked funnier than captions which are funny by themselves. In addition, we can see that funny captions' most distinct word (joke word, if exists) are less distinct than these of their counterparts. This supports the findings of [17], discussed above.

The table also shows that funny captions have fewer joke words in every quarter, but this can be explained by the previous point. If funny captions have less jokes, it makes sense that they have less jokes per quarter as well. When restricting the test to pairs where both captions contain joke words, the only location feature that remains significant is the second quarter, achieving a score of 0.42***. This seems to indicate that less funny captions tend to reveal the joke too early, potentially not providing as much of a build-up or generation of tension, highlighted as important in psychological theories of humor.

| Feature | %Funnier is higher |
|---|---|
| Perplexity (1-gram) | 0.48 |
| Perplexity (2-gram) | 0.51* |
| Perplexity (3-gram) | 0.52* |
| Perplexity (4-gram) | 0.52* |
| POS Perplexity (1-gram) | 0.47*** |
| POS Perplexity (2-gram) | 0.5 |
| POS Perplexity (3-gram) | 0.5 |
| POS Perplexity (4-gram) | 0.51 |
| Indefinite articles | 0.52 |
| 3rd Person | 0.45** |
| Min Perplexity | 0.46***** |
| Length of Joke Phrase | 0.42****** |
| Location (quarters) | 0.46, 0.41*****, 0.46*, 0.43**** |
| Proper Nouns | 0.35****** |
| Sentiment | 0.49 |
| Readability | 0.52, 0.51 |
| Similarities (context, anomaly, maxdiff, avgdiff) | 0.48*, 0.48, 0.47**, 0.47** |

**Table 3: Different jokes: Percentage of caption pairs in which a feature has higher numerical value in the funnier caption. Significance according to a two-sided Wilcoxon signed rank test is indicated using *-notation (*p≤0.05, **p≤0.005, Holm-Bonferroni correction)**

## 3.4 Prediction Task

As in Section 2.4, we trained a random forest classifier on the data. Using bag-of-words features alone, the random-forest classifier achieved 55% (10-fold cross validation). The important word list was similar, with a few more verbs (mean, see, told). Using all the features, however, the performance improved to 64%.

As before, we evaluate the importance of features using the Gini importance. The highest-scoring features included perplexity features (both lexical and POS), similarities to context and anomalies, sentiment, readability and proper nouns. Again, no bag-of-words features made it into the top 50 important features.

We note that some of our features (e.g., 3rd person, indefinite articles) do not appear to contribute significantly to the classifier's accuracy. This may be due to the fact that there are many different elements of style, and the contribution of each tends to be quite subtle. However, they all tend to appear indirectly in the distribution of the higher-level features (e.g., the ones derived from language models). Importantly,

| | |
|---|---|
| - Just listen to that baby purr. | - The previous owner played Frisbee. |
| - And it purrs like a kitten. | - This is the Mordor package. |
| - This model features the latest advances in hybrid technology. | - Yes, GMM! Genetically Modified Mechanism! |
| -.and it won't cost you arm and a leg. | - We call him Fighting McQueen. |
| - Highly recommended when you're fighting for a parking space. | - This ain't your father's Buick Wildcat. |
| - This baby purrs like a cat. | - No gas. No nightly recharging! Just...um... |
| - This baby growls like a bear. | - The Paleo,great mileage,runs on road kill. |
| - Best to not kick the tires. | - The previous owner was Lon Cheney |
| - Just don't kick the tires. | - Stephen King's loss is your gain. |
| - Just don't ask for leather. | - And it's a bear to drive!p |
| - You'll save thousands on tires. | - ...Yes! Very low emissions and usually only after dinner. |
| - Welll, no, I can't promise that it won't roll over. | - It runs on a 100% fuel efficient Paleo Diet. |
| - You sure you don't want to look under the hood? | - We also have the American model with shaved legs. |
| - This is a true hybrid. | - Sure, take a test run. But don't kick the tires. |
| - Oh, I almost forgot...It also comes with full moon insurance. | - We call her a 2015, but technically she's a 14105. |

**Figure 6: Tournament results: High-ranking (left) and low-ranking (right) captions for the cartoon in Figure 1. We only depict captions of length 5-10, and some near-duplicates are removed for the sake of presentation.**

classifiers excel at making use of those indirect cues. Thus, features that improve prediction might not necessarily align with features that provide insights.

## 4. FINDING THE BEST CAPTIONS

We now have a classifier that can compare pairs of captions. However, our greater goal is to reduce the workload of the contest judges. At present, judges have to manually sift through thousands of captions. We next explore whether the classifier can be used to filter the submissions down to a more manageable size, allowing the judges to find the best captions faster.

We decided to use our classifier in a *tournament* algorithm, pairing captions against each other to find the best ones. Since we were interested in finding a set of the best captions (as opposed to the top one), we employed a non-elimination tournament algorithm. We chose the Swiss-system tournament method. In this method, competitors are paired randomly in the first round; in the next rounds, they are paired based on their performance so far. We group competitors by total points and add edges between players who have not been matched yet. This ensures that no two competitors face each other more than once. We then use the Blossom algorithm to compute a maximal matching.

For each cartoon, we trained a classifier using data on all other cartoons. This is meant to simulate a new contest. Whenever two captions were paired, we computed the classifier's prediction. Since many pairs of captions are incomparable (people do not reach a wide agreement on ranking), we declared a tie when the classifier's confidence was low. A victory was worth 3 points, and a tie was worth 1 point.

Figure 6 shows some of the top-ranking (left) and bottom-ranking (right) captions for the cartoon in Figure 1. Many of the captions that were ranked at the very bottom were very long (15-42 words each); to make the comparison between the captions easy, we only depict captions of length 5-10. For the sake of the presentation, we remove near-duplicates (for example, many variations on "You should hear the engine purr" received a very similar score). We also note that many of the lowest-ranked captions contain proper nouns.

To evaluate our methods, we obtained the 9-10 shortlisted captions chosen by the New Yorker editor for each contest. We ran a tournament for each cartoon, and recorded the rank of the shortlisted captions. On average, all of the shortlisted captions appeared in the top 55.8% of their respective tournament, with the best tournament achieving 48.9% and the worst achieving 71.4%.

Figure 7 shows the cumulative results across all the contests: The tournaments' top-ranked $x$% of the captions contain $y$% of all of the shortlist captions. For example, 37% of the shortlisted captions were ranked within their tournament's top 10%. 80% of them appear before the 50% mark, and all of them are covered around the 72% mark. In a way, this is the empirical tradeoff that our algorithm offers the contest judge.

We note that for many of the shortlisted captions that were ranked low, another caption from the same cluster ap-

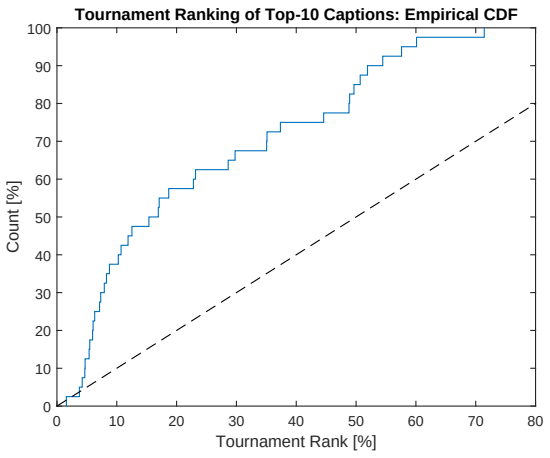



**Figure 7: Rank of shortlisted captions in the tournament. The tournament's top-ranked $x$% of the captions contain $y$% of the shortlist captions. For example, 37% of the shortlisted captions were ranked within the tournament's top 10%.**

peared higher in the ranking. For example, "The last guy got flushed", from the editor's shortlist, was ranked 248; "They flushed the last guy" was ranked 85. This suggests that combining tournament ranking with clustering might improve the results and provide more benefit to the judges.

The tournament results are particularly interesting since the classifier was trained on Mechanical Turk workers, whose sense of humor is likely different from the New Yorker audience (or its proxy, the contest organizer). For example, in Patrick House's article about his winning strategy [15], he advises to write what he calls "theory of mind" captions – "higher-order jokes easily distinguished from the simian puns and visual gags that litter the likes of MAD Magazine". However, when we asked the crowdworkers to justify their choice of funniest caption, "pun" (along with "punny", "play on words", "word play") was a very popular explanation.

**Limitations.** Looking at the editor's shortlisted captions that were ranked the lowest can reveal some of the limitations of our algorithm. For example, consider the caption "Looks like you just missed Him" for the cartoon in Figure 3. The joke only works because of the capitalization of "Him" (often used when referring to God). However, this capitalization is completely lost on some of our case-insensitive algorithms, in particular the word embeddings.

Similarly, the caption "I liked the old NSA better" (said by two office workers sitting inside a fishbowl) is penalized for using a proper noun. In this case, the algorithm is oblivious to recent high-profile news events and the societal connotations that this specific proper noun carries.

# 5. RELATED WORK

Humor is well-studied in fields such as linguistics and psychology. However, there have been only a limited number of contributions towards computational humor prototypes. In this section, we briefly review research contributions from all of these fields.

**Psychology.** Several studies in psychology have focused specifically on cartoons. For example, creativity was found to be linked to the ability to generate cartoon captions [7].

Jones et al. demonstrated the importance of a relationship between picture and caption [18]. Subjects viewed picture and caption either simultaneously or sequentially (picture appearing five seconds before the caption, or vice versa). More vigorous changes in heart rate where detected when the elements were shown in sequence versus simultaneously.

Cartoons have also been used to test the incongruity theory. In one study, the original, incongruity-removed and resolution-removed cartoons were shown to elementary-school children in order to assess the humor-inducing effects of incongruity and resolution [28].

The work of Jones [17] is perhaps most related to our work. This study analyzed the components of cartoon humor and determined the effects of caption, picture, and the interaction between them on the humor rating of the whole cartoon. Their results indicated that the funnier the picture and the *less funny* the caption, the higher the cartoon was ranked, which is somewhat supported by our results. These findings also support the incongruity theory for cartoons.

**Linguistics.** There has been considerable research on theories of humor and linguistics of humor [2]. Within linguistics, most studies attempt to develop symbolic relationships between linguistic units, examine surface attributes of jokes and attempt to classify them [16, 9].

Raskin's seminal work [26] presented a framework to formulate incongruity theory. The framework uses scripts, which are structured configurations of knowledge about some situation or activity. Scripts were later developed into the General Theory of Verbal Humor (GTVH) [4]. This theory proposes a hierarchy of types of knowledge that is used when composing of a joke. GTVH is interesting, but is not quite ready to be automated. Many of its basic constructs are informal, and rely on the experimenter's intuitions.

**Computer Science.** The JAPE punning-riddle generator [5] is one of the first attempts of computational humor. The HAHAcronym project [29] focused on automatically generating humorous acronyms. Humor was achieved by mixing terms from different domains, exploiting incongruity theory.

On the side of humor understanding, notable works focused on knock-knock jokes [30] and one-liners [23]. Recently, Tsur et al. tackled the related problem of identifying sarcasm [32]. In contrast to our work, they have tried to identify common patterns, e.g., "does not _ much about _ _ or". To the best of our knowledge, the problem of automatically identifying funny cartoon captions is novel.

# 6. CONCLUSIONS AND FUTURE WORK

We investigated the challenge of learning to recognize the degree of humor perceived for combinations of captions and cartoons. We extracted a useful set of features from linguistic properties of captions and the interplay between captions and pictures. We harnessed a large corpus of crowdsourced cartoon captions that were submitted to a contest hosted by the New Yorker. We developed a classifier that could pick the funnier of two captions 64% of the time, and used it to find the best captions, significantly reducing the load on the cartoon contest's judges.

We believe that the framework developed in this paper can serve as a basis for further research in computational humor. Future directions of work include more detailed analysis of the context and anomalies represented in the cartoons, as well as their influence on setting up tension, questions, or confusion. We seek to better understand the nature of humorous explanations that connect the disparate frames in a surprising or interesting manner. We believe there is much to be done with extending our geometric approach to modeling distances between multiple frames.

We also see value in applying and leveraging methods for automatically detecting core contextual and anomalous frames in a cartoon. We are interested in harnessing gaze-tracking devices to investigate how people explore visual components of a cartoon in the absence of captions, as well as with alternate captions.

We are also interested in humor generation for captions and cartoons. In principle, we could use our classifier to try and improve an existing caption. We could identify weaknesses of the caption, and suggest improvements (for example, replacing a word by a simpler synonym). Furthermore, it may be worthwhile to explore the generative process employed by people in coming up with captions. For example, we are interested in understanding the influence of the visual salience of the core context and anomalies of a scene on the focus and linguistic structure of a joke.

Finally, humor is a matter of personal style and taste. Thus, it might be interesting to explore opportunities to personalize the classifier. For example, we can treat each crowdworker as a (partial) order on the captions. In order to personalize a classifier, we could ask a person to provide several rankings, and then use a rank-based collaborative filtering algorithm (similar to [21]) or a variant of Kemeny distance to find people with a similar taste.

Automated generation and recognition of humor could be harnessed to modulate attention, engagement, and retention of concepts, and thus has numerous interesting applications, including use in education, health, engagement, and advertising. Beyond applications, we believe that pursuing information-theoretic models of humor could reveal new insights about one of the most fascinating human behaviors.

**Acknowledgments.** We thank Ben Schwartz for his accurate portrayal of the authors.

# 7. REFERENCES

[1] Tesla stock moves on april fools' joke, http://blogs.wsj.com/moneybeat/2015/04/01/tesla-stock-moves-on-april-fools-joke, 2015.

[2] S. Attardo. *Linguistic Theories of Humor*. Approaches to Semiotics. Mouton de Gruyter, 1994.

[3] S. Attardo. A primer for the linguistics of humor. *The Primer of Humor Research, Berlin & New York: Mouton de Gruyter*, pages 101–155, 2008.

[4] S. Attardo and V. Raskin. Script theory revis (it) ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*, 1991.

[5] K. Binsted and G. Ritchie. An implemented model of punning riddles. In *AAAI'94*, 1994.

[6] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[7] D. M. Brodzinsky and J. Rubien. Humor production as a function of sex of subject, creativity, and cartoon content. *Journal of consulting and Clinical Psychology*, 44(4):597, 1976.

[8] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint 1312.3005*, 2013.

[9] C. Curcó. The implicit expression of attitudes, mutual manifestness and verbal humour. *UCL Working Papers in Linguistics*, 8:89–99, 1996.

[10] C. Danescu-Niculescu-Mizil, J. Cheng, J. Kleinberg, and L. Lee. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL '12, 2012.

[11] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[12] R. Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[13] E. Gabrilovich, M. Ringgaard, and A. Subramanya. Facc1: Freebase annotation of clueweb corpora, 2013.

[14] D. Hofstadter and L. Gabora. Synopsis of the workshop on humor and cognition. *arXiv preprint arXiv:1310.1676*, 2013.

[15] P. House. How to win the new yorker cartoon caption contest. Slate, 2008.

[16] M. Jodłowiec. What's in the punchline? *Relevant Worlds: Current Perspectives on Language, Translation and Relevance Theory. Newcastle: Cambridge Scholars Publishing*, pages 67–86, 2008.

[17] J. M. Jones, G. A. Fine, and R. G. Brust. Interaction effects of picture and caption on humor ratings of cartoons. *The Journal of Social Psychology*, 108(2):193–198, 1979.

[18] J. M. Jones and P. E. Harris. Psychophysiological correlates of cartoon humor appreciation. In *Proceedings of the Annual Convention of the American Psychological Association*. American Psychological Association, 1971.

[19] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[20] B. King, R. Jha, D. R. Radev, and R. Mankoff. Random walk factoid annotation for collective discourse. In *ACL*, pages 249–254, 2013.

[21] N. N. Liu and Q. Yang. Eigenrank: A ranking-oriented approach to collaborative filtering. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 83–90, New York, NY, USA, 2008. ACM.

[22] A. P. McGraw and C. Warren. Benign violations making immoral behavior funny. *Psychological Science*, 21(8):1141–1149, 2010.

[23] R. Mihalcea and S. Pulman. Characterizing humour: An exploration of features in humorous texts. In *Computational Linguistics and Intelligent Text Processing*, pages 337–347. Springer, 2007.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013.

[25] E. Oring. *Engaging humor*. University of Illinois Press, 2003.

[26] V. Raskin. *Semantic mechanisms of humor*, volume 24. Springer Science & Business Media, 1985.

[27] R. Senter and E. Smith. Automated readability index. Technical report, DTIC Document, 1967.

[28] T. R. Shultz. The role of incongruity and resolution in children's appreciation of cartoon humor. *Journal of Experimental Child Psychology*, 13(3):456–477, 1972.

[29] O. Stock and C. Strapparava. Hahacronym: A computational humor system. In *Proceedings of the Association for Computational Linguistics*. ACL, 2005.

[30] J. Taylor and L. Mazlack. Computationally recognizing wordplay in jokes. *Proceedings of CogSci 2004*, 2004.

[31] L. L. Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

[32] O. Tsur, D. Davidov, and A. Rappoport. Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *ICWSM*, 2010.