A computational model of linguistic humor

Justine T. Kao [*], **Roger Levy** [†] **and Noah D. Goodman** [*]

[*]Stanford University, and [†]University of California, San Diego

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Humor plays an essential role in human interactions. However, its precise nature remains elusive. While research in natural language understanding has made significant advancements in recent years, there has been little direct integration of humor research with computational models of general language understanding. In this paper, we propose two information-theoretic measures of humor–ambiguity and distinctiveness–inspired by humor theories and directly derived from a simple model of sentence processing. We then test these measures on a set of puns and non-pun sentences and show that they correlate significantly with human judgments of funniness. Our model is one of the first to integrate general linguistic knowledge, speaker intent, and humor theory to model humor computationally. We present it as a framework for applying models of language processing to understand higher-level linguistic and cognitive phenomena.

Linguistic humor | Language understanding | Computational modeling

Abbreviations: IR, Incongruity Resolution

Introduction

Imagine living a day without humor. From friendly exchanges with an affable stranger to sidesplitting laughter among close friends, our everyday experiences clearly show that humor plays an essential role in human interactions. Many studies suggest that humans are naturally attracted to and can often significantly benefit from humorous stimuli [3-8]. Here we develop a computational model to examine this ubiquitous and fundamental phenomenon, with the aim of shedding light on the conditions in which the mind perceives humor in language.

Researchers in artificial intelligence have argued that given the importance of humor in human communication, computers need to generate and detect humor in order to interact with humans more effectively mihalcea2006learning. However, most work in computational humor has focused either on joke-specific templates and schemata binsted1996machine, kiddon2011s or surface linguistic features that predict humorous intent mihalcea2006learning, semantic2010. The former type of studies is restricted to identifying jokes with a very specific format and structure, and the latter type falls short of testing or building upon more general theories of humor. Our work moves beyond these approaches and directly utilizes a model of sentence comprehension to derive theory-driven measures of humor.

Incongruity, defined as incompatible and often schema-violating interpretations of a stimulus, has received the most attention as a compelling requisite of humor [cite]. As Veale (2004) states, Of the few sweeping generalizations one can make about humor that are neither controversial or trivially false, one is surely that humor is a phenomenon that relies on incongruity. However, other scholars argue that incongruity alone is insufficient, since incongruous situations may simply appear senseless or dissonant instead of humorous [cite]. These scholars propose a two-stage model of humor termed the incongruity-resolution (IR) model, in which resolution is commonly understood as discovery of a cognitive rule that reconciles the incongruous parts of a situation [cite].

While informal models of humor have provided a useful framework for analyzing necessary and sufficient conditions of humor, the lack of computational rigor makes it difficult to operationalize and empirically evaluate the role of different cognitive factors in the perception of humor. Within the IR model, definitions of incongruity and resolution are often ambiguous and leave much room for disparate interpretations across scholars [cite]. In this paper, we use a computational model of language to formalize incongruity and resolution in order to empirically evaluate their relationships to linguistic humor. While we aim to develop a model that encompasses a broad range of humorous stimuli, a critical challenge lies in the fact that complex cognitive phenomena like humor rely on rich commonsense knowledge and discourse understanding, which are challenging topics and largely unsolved in both artificial intelligence and cognitive science. To somewhat limit the scope of our task and work within the constraints of formal representations of meaning, we focus on applying formalizations of incongruity and resolution to a subset of linguistic humor: puns. In particular, we focus on phonetic punspuns containing words that sound identical or similar to other words in the English languagebecause the space of possible interpretations and meaning representations of a phonetic pun is relatively constrained and well defined.

An example helps to illustrate: The magician got so mad he pulled his hare out.

This sentence allows for two interpretations:

(a) The magician got so mad he performed the trick of pulling a rabbit out of his hat.
(b) The magician got so mad he (idiomatically) pulled out the hair on his head.

If the comprehender interprets the word hare as itself, he will arrive at interpretation (a); if he interprets the word as its homophone hair, he will arrive at interpretation (b). The sentence-level differences between interpretations (a) and (b) can thus be approximated by the two interpretations of the observed word hare. In general, distinct interpretations of a

Table 1. Examples of identical homophone sentences

Sentence Type	Example
Pun	A dentist has to tell a patient the whole tooth.
Non-pun	A dentist examines one tooth at a time.
Non-pun	She always speaks the truth.

Reserved for Publication Footnotes

phonetic pun hinges on one phonetically ambiguous word, allowing the two lexical forms of the ambiguous word to stand in for competing interpretations of the entire sentence. This allows us to tackle humor without having to solve the problem of representing meaning in complex sentences or discourse.

Sentence Type	Example
Pun	The magician was so mad he pulled is hare out.
Non-pun	The hare ran rapidly across the field.
Non-pun	Some people have lots of hair on their heads.

Critically, even though the example we gave was a written pun and the reader sees the word "hare" explicitly on the page, the "hair" interpretation is still present and even salient in the context of the sentence. The possibility of interpretations that are distinct from the observed sentence itself is captured by noisy channel models of sentence processing, which posit that language comprehension is a rational process that incorporates uncertainty about surface input to arrive at sentence-level interpretations that are globally coherent [cite]. Comprehenders can thus consider multiple word-level interpretations to arrive at more than one interpretation of a sentence, each coherent but potentially incongruous with each other. The notion of incongruity fits naturally into a noisy channel model of sentence comprehension. This intuition also corresponds with that of writer and philosopher Henri Bergson, who defined a pun as a sentence or utterance in which two ideas are expressed, and we are confronted with only one series of words."

While our approach is novel in its quantitative rigor, it is directly informed by incongruity-resolution models of humor. Since incongruity and resolution are properties of the interpretations derived from a sentence, we first describe a probabilistic model of sentence interpretation. Our model aims to infer the topic of a sentence (a coarse representation of its meaning) from the observed words. Unlike previous such models, however, we take a noisy channel approach, assuming that the comprehender maintains uncertainty over which words reflect the sentence topic and which are noise.

Assume our sentence is composed of a vector of content words $\vec{w} = \{w_1, \ldots, w_i, h, w_{i+1}, \ldots, w_n\}$, including a phonetically ambiguous word h. We will use a simple generative model for $\vec{w}$ (see Figure). Our model is motivated by the important roles that both semantic priming and the sequential structure of language play in lexical disambiguation during sentence processing [cite]. Given the latent sentence topic m, each word is generated independently by first deciding if it reflects the topic (the indicator variable f_i). If so it is sampled based on semantic relevance to m; if not it is sampled from an n-gram model that takes into account the immediately preceding words. We thus view the sentence as a mixture of topical and non-topical words. Similar approaches have been used in generative models of language to account for words that provide non-semantic information, such as topic models that incorporate syntax [cite].

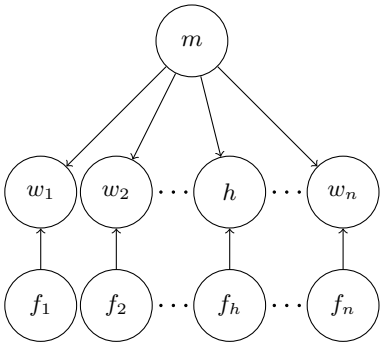

Fig. 1. Generative model of a sentence. Each word w_i is generated based on the sentence topic m if the indicator variable f_i puts it in semantic focus; otherwise it is generated as noise (from a unigram distribution).

We make the simplifying assumption that the plausible candidate topics m of the sentence correspond to the potential interpretations of the homophone word h, which are constrained by phonetic similarity to two alternatives, m_1 and m_2. For example, in the magician pun described above, h is the phonetically ambiguous target word "hare," and m_1 and m_2 are the candidate interpretations *hare* and *hair*. The two potential topics of the sentence can be identified by the two interpretations *hare* and *hair*. This assumption reduces the ill-defined space of sentence meanings to the simple proxy of alternate spellings for phonetically ambiguous words.

Using the above generative model, we can infer the joint probability distribution $P(m, \vec{f}|\vec{w})$ of the sentence topic m and the indicator variables $\vec{f}$ that determine whether each word is in semantic focus. This distribution can be factorized into:

$$P(m, \vec{f}|\vec{w}) = P(m|\vec{w})P(\vec{f}|m, \vec{w}) \quad [1]$$

The two terms on the right-hand side are the basis for our derivations of measures for incongruity and resolution, respectively. Incongruity means the presence of two similarly likely interpretations, which can be quantified as a summary of the binomial distribution $P(m|\vec{w})$. If the entropy of this distribution is low, then only one meaning is likely; if the entropy is high, then both meanings are similarly likely. Under the assumption that the two candidate meanings are sufficiently different from each other, high entropy in the meaning distribution suggests coexistence of two incompatible interpretations, which directly characterizes incongruity.

Resolution measures the degree to which two interpretations are supported by distinct parts of the sentence. We represent this as the divergence between sets of words that are in semantic focus given the two values of m, which can be quantified as a summary of the distribution $P(\vec{f}|m, \vec{w})$. We use a Kullback-Leibler divergence score $D_{KL}(F_1||F_2) + D_{KL}(F_2||F_1)$ to measure the distance between F_1 and F_2. This score measures how distinct" the semantic focus sets are given m_1 and m_2. A low KL score would indicate that meanings m_1 and m_2 are supported by similar subsets of the sentence; a high KL score would indicate that m_1 and m_2 are both strongly supported by distinct subsets of the sentence. Together, these two measures constitute our formalization of humor as informed by a two-stage incongruity-resolution model.

Results

We evaluate the contribution of each of our quantitative measures of incongruity and resolution to humor by correlating the measures with humans judgments of humor in two separate corpora of phonetically ambiguous sentences. We first evaluate our measures on 195 sentences, in which the two candidate meanings of the phonetically ambiguous word sound identical to each other. Of these sentences, 65 are identical homophone puns and 130 are non-pun control sentences that match the puns in containing the same phonetically ambiguous words. To confirm that our measures generalize to broader types of ins, we also evaluate on 240 sentences in which the two candidate meanings of the phonetically ambiguous word sound

similar but not identical to each other. Of these sentences, 80 are near homophone puns and 160 are non-pun control sentences.

Table 2. Examples of near homophone sentences

	Estimate	Std. Error	p value
Intercept	-0.830	0.1070	< 0.0001
Ambiguity	1.899	0.212	< 0.0001
Distinctiveness	0.568	0.082	< 0.0001

Following the derivations and using the relatedness measures described in the Appendix, we computed an incongruity and distinctiveness value for each of the 435 sentences. As predicted, incongruity differs significantly across sentence types ($F(1, 433) = 108.4, p < 0.0001$) and correlates significantly with human ratings of funniness across all pun and non-pun sentences ($r = 0.42, p < 0.0001$). However, incongruity does not correlate with human ratings of funniness within the 145 pun sentences ($r = 0.03, p > 0.05$).

On the other hand, distinctiveness differs marginally significantly across sentence types ($F(1, 433) = 47.1, p < 0.01$) and correlates significantly with human ratings of funniness across all 435 sentences, although to a lesser extent than incongruity scores ($r = 0.35, p < 0.001$). Importantly, distinctiveness ratings correlate significantly with human ratings of funniness within the pun sentences only ($r = 0.28, p < 0.001$). This suggests that while high ambiguity distinguishes puns from non-puns, distinctiveness of the supporting context for each meaning is needed to separate very funny puns from mediocre ones.

A linear regression showed that both incongruity and resolution are significant predictors of funniness. Together, the two predictors capture a modest but significant amount of the variance in funniness ratings ($F(2, 432) = 76.79, R^2 = 0.26, p < 0.0001$; see Table 3). Using both incongruity and resolution as dimensions that formalize humor, we can distinguish among pun and non-pun sentences, as shown in Figure 2. Figure 2 shows the standard error ellipses for the two sentence types in the two-dimensional space of incongruity and resolution. Although there is a fair amount of noise in the predictors (likely due to simplifying assumptions, the need to use empirical measures of relatedness, and the inherent complexity of humor) we see that pun sentences tend to cluster at a space with higher incongruity and resolution, while non-puns score significantly lower on both incongruity and resolution.

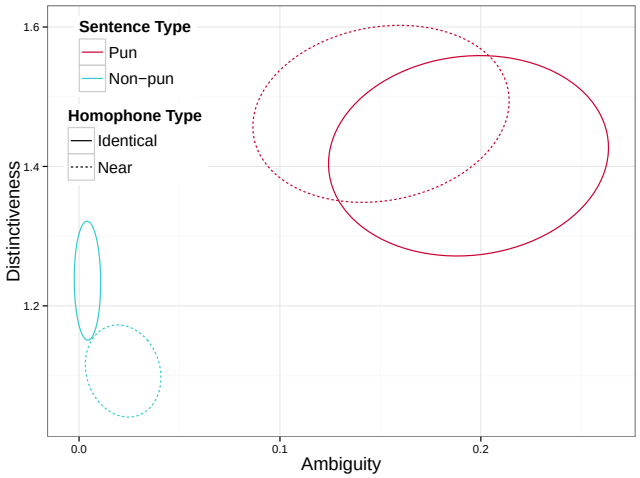

Fig. 2. Standard error ellipses of ambiguity and distinctiveness across sentence types. Puns score higher on ambiguity and distinctiveness; non-puns have low ambiguity and distinctiveness.

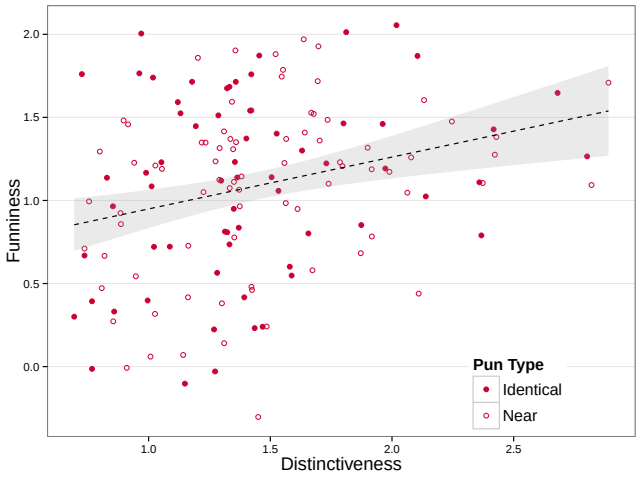

Fig. 3. Figure caption

Simulations.

Table 3. Regression coefficients using ambiguity and distinctiveness to predict funniness ratings

$m1$	$m2$	Type	Sentence and Semantic Focus Sets	Amb.	Disj.	Funniness
hare	hair	Pun	The **magician** got so mad he **pulled** his **hare** out.	0.570	3.405	1.714
		De-pun	The professor got so mad he **pulled** his **hare** out.	0.575	2.698	0.328
		Non-pun	The **hare ran rapidly** through the **fields**.	0.055	2.791	-0.400
		Non-pun	Most **people** have **lots** of **hair** on their **heads**.	$2.76E^{-5}$	3.920	-0.343
tiers	tears	Pun	It was an **emotional wedding**. Even the **cake** was in **tiers.**	0.333	3.424	1.541
		De-pun	It was an **emotional wedding**. Even the **mother-in-law** was in **tiers**.	0.693	2.916	0.057
		Non-pun	**Boxes** are **stacked** in **tiers** in the warehouse.	0.018	3.203	-0.560
		Non-pun	**Tears** ran down her **cheeks** as she watched a **sad movie**.	$1.73E^{-5}$	4.397	-0.569

Simulation 1

Simulation 2

Real Data.

Discussion

We believe our work represents a step towards developing models of language that can capture rich social and linguistic meaning. From the perspective of language understanding, such phenomena can serve as probes for developing models of language that account for the subtleties of linguistic behavior. From the perspective of humor research, such computational models allow for formalizations that can help empirically validate and refine existing theories. We hope that our work contributes to research in humor theory, computational humor, and language understanding, with the aim to one day understand what makes us laugh and build robots that appreciate the wonders of word play.

Materials and Methods

Our first dataset consists of 40 identical homophone pun sentences from a website called "Pun of the Day" (). Puns were selected such that the ambiguous item in each pun is a single phonetically ambiguous word. To obtain more identical homophone pun items, a research assistant generated an additional 25 pun sentences based on a separate list of homophone words. We then selected 130 corresponding non-pun sentences from an online version of Heinle's Newbury House Dictionary of American English (). We chose sample sentences included in the definition of the homophone word. 65 of the sentences contain the ambiguous words from the pun sentences, and 65 of them contain the alternative homophones. This design ensured that puns and non-pun sentences contain the same set of phonetically ambiguous words. Table 1 shows example sentences from each category. Our second dataset consists of 80 near homophone pun sentences from the same pun website, as well as 160 corresponding near homophone non-pun sentences.

We obtained funniness ratings for the two datasets. The195 identical homophone sentences were rated by 100 subjects on Amazon's Mechanical Turk. Each subject read roughly 60 sentences in random order, counterbalanced for the sentence types, and rated each sentence on funniness and correctness. The average split-half correlation of funniness ratings was 0.83. Figure **??** shows the average funniness ratings of puns and non-pun sentences. Pun sentences are rated as significantly funnier than non-pun sentences ($F(2, 232) = 415.3, p < 0.0001$).

As described in the model section, computing ambiguity and distinctiveness measures requires the prior probabilities of meanings $P(m)$ (approximated as the unigram probabilities of the words that denote the meanings), the prior probabilities of words $P(w)$, and the conditional probabilities of each word in the sentence given a meaning $P(w|m)$. While we computed $P(w)$ and $P(m)$ directly from the Google Web unigram corpus, $P(w|m)$ is difficult to obtain through traditional topic models trained on corpora due to data sparsity. Since each meaning we consider has a single word as proxy, we may approximate $P(w|m)$ using an empirical measure of the semantic relatedness between w and m, denoted $R(c, m)$. We use $R(c, m)$ as a proxy for point wise mutual information between c and m, defined as follows:

$$R(w, m) = \log \frac{P(w, m)}{P(w)P(m)} = \log P(w|m) - \log P(w) \quad [2]$$

We assume that human ratings of relatedness between two words $R'(w, m)$ approximate true relatedness up to an additive constant z. With the proper substitutions and transformations,

$$P(w|m) = e^{R'(w,m)+z} P(w) \quad [3]$$

To obtain $R'(w, m)$ for each of the words w in the stimuli sentences, we recruited 200 subjects on Amazon's Mechanical Turk to rate distinct word pairs on their semantic relatedness. Since it is difficult to obtain the relatedness rating of a word with itself, we used a free parameter r and fit it to data. Function words were removed from each of the sentences in our dataset, and the remaining words were paired with each of the interpretations of the homophone sequence (e.g., for the pun in Table 1, "magician" and "hare" is a legitimate word pair, as well as "magician" and "hair"). This resulted in 1460 distinct word pairs. Each subject saw 146 pairs of words in random order and were asked to rate how related each word pair is using a scale from 1 to 10. The average split-half correlation of the relatedness ratings was 0.916, indicating that semantic relatedness was a reliable measure.

Definition 1.

Theorem 1.

Appendix

An appendix without a title.

Appendix: Appendix title

An appendix with a title.

ACKNOWLEDGMENTS. This work was partially supported by a grant from the Spanish Ministry of Science and Technology.

1. M. Belkin and P. Niyogi, Using manifold structure for partially labelled classification, Advances in NIPS, 15 (2003).
2. P. Bérard, G. Besson, and S. Gallot, Embedding Riemannian manifolds by their heat kernel, Geom. and Fun. Anal., 4 (1994), pp. 374–398.
3. R.R. Coifman and S. Lafon, Diffusion maps, Appl. Comp. Harm. Anal., 21 (2006), pp. 5–30.
4. R.R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data. Part I: Diffusion maps, Proc. of Nat. Acad. Sci., (2005), pp. 7426–7431.
5. P. Das, M. Moll, H. Stamati, L. Kavraki, and C. Clementi, Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction, P.N.A.S., 103 (2006), pp. 9885–9890.
6. D. Donoho and C. Grimes, Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data, Proceedings of the National Academy of Sciences, 100 (2003), pp. 5591–5596.
7. D. L. Donoho and C. Grimes, When does isomap recover natural parameterization of families of articulated images?, Tech. Report Tech. Rep. 2002-27, Department of Statistics, Stanford University, August 2002.
8. M. Grüter and K.-O. Widman, The Green function for uniformly elliptic equations, Man. Math., 37 (1982), pp. 303–342.
9. R. Hempel, L. Seco, and B. Simon, The essential spectrum of neumann laplacians on some bounded singular domains, 1991.
10. Kadison, R. V. and Singer, I. M. (1959) Extensions of pure states, Amer. J. Math. 81, 383-400.
11. Anderson, J. (1981) A conjecture concerning the pure states of $B(H)$ and a related theorem. in Topics in Modern Operator Theory, Birkhaüser, pp. 27-43.
12. Anderson, J. (1979) Extreme points in sets of positive linear maps on $B(H)$. J. Funct. Anal. 31, 195-217.
13. Anderson, J. (1979) Pathology in the Calkin algebra. J. Operator Theory 2, 159-167.
14. Johnson, B. E. and Parrott, S. K. (1972) Operators commuting with a von Neumann algebra modulo the set of compact operators. J. Funct. Anal. 11, 39-61.
15. Akemann, C. and Weaver, N. (2004) Consistency of a counterexample to Naimark's problem. Proc. Nat. Acad. Sci. USA 101, 7522-7525.
16. J. Tenenbaum, V. de Silva, and J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science, 290 (2000), pp. 2319–2323.
17. Z. Zhang and H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignement, Tech. Report CSE-02-019, Department of computer science and engineering, Pennsylvania State University, 2002.