Distribution-Based Measures of Surprise for Creative Language: Experiments with Humor and Metaphor

Razvan C. Bunescu
Department of Computer Science
University of North Carolina at Charlotte
Charlotte, NC 28223
razvan.bunescu@uncc.edu

Oseremen O. Uduehi
School of EECS
Ohio University
Athens, OH 45701
ou380517@ohio.edu

Abstract

Novelty or surprise is a fundamental attribute of creative output. As such, we postulate that a writer's creative use of language leads to word choices and, more importantly, corresponding semantic structures that are unexpected for the reader. In this paper we investigate measures of surprise that rely solely on word distributions computed by language models and show empirically that creative language such as humor and metaphor is strongly correlated with surprise. Surprisingly at first, information content is observed to be at least as good a predictor of creative language as any of the surprise measures investigated. However, the best prediction performance is obtained when information and surprise measures are combined, showing that surprise measures capture an aspect of creative language that goes beyond information content.

1 Introduction

Language is used primarily as a means for communicating information. It is thus appropriate that information theory (Shannon, 1948) has provided the foundation for numerous studies into properties of natural language, as in (Shannon, 1951; Hale, 2001; Piantadosi et al., 2011; Gibson, 2019), among many others. Under the information theory framework, a communication channel is posited between the speaker and the listener, and correspondingly the goal of the speaker is to employ the channel as efficiently as possible while also minimizing the risk of miscommunication. Maximizing the use of the communication channel is achieved when speakers choose their words such that their information rate is close to the channel capacity, which can be seen as determining speakers to construct utterances such that information is spread uniformly across them. This is known as the Uniform Information Density (UID) hypothesis (Fenk and Fenk-Oczlon, 1980; Jaeger and Levy, 2006), operationalized as a tendency for regression towards the mean

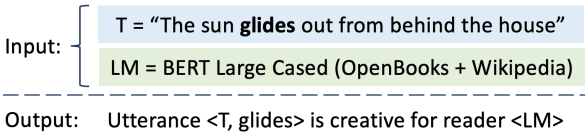

Figure 1: Creative language detection requires as input not only the Text (T), but also the Reader (LM).

information content across the language (Meister et al., 2021). The UID hypothesis can explain a variety of linguistic phenomena, such as the optional omission of syntactic relativizers (Jaeger and Levy, 2006), or the shortened phonetic duration of highly predictable language units (Aylett and Turk, 2004). UID has also been construed to imply that speakers avoid producing words with an *information content*[1] that is too high or too low (Meister et al., 2022) relative to the expected information rate of the channel, or the entire language. While this holds true for most communicative uses of language, there are at least two types of situations when words have an information content much higher than expected, as illustrated in Figure 1.

First, there is the case when the listener has no clear expectation of what the speaker will utter next, such as when introducing a new discourse entity through a definite or indefinite article, especially at the beginning of a story when not much context is available. In this case, the next word distribution has a high entropy, all words have a relatively low probability, hence high information content. The word 'sun' in the sentence[2] shown in Figure 1 is in this category. Second, there are situations when language is used in creative ways, when speakers deliberately produce words or phrases that are interesting or unexpected, often with the purpose of inducing particular kinds of emotion in the listener, as is the case with the word 'glides' in Figure 1. In this paper, we aim to characterize such creative use of language solely through distribution-based

[1] Computed as negative log of word probability $-\log p(x)$.
[2] First line in a poem by Tomas Tranströmer.

Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 68 - 78

measures that are designed to discriminate creative language from normal language. In both situations discussed above the information content is high, therefore, at least theoretically, information content alone is not sufficient to discriminate between the two. As such, we propose that *surprise* be used as the main discriminating factor. We emphasize that determining whether an input text exhibits creativity or surprise requires specifying a *reference reader*, as shown in the example from Figure 1, which distinguishes the task explored in this paper from related tasks such as humor detection or figurative language classification, where novelty with respect to a reference reader is not a concern.

2 Definitions and Measures of Surprise

The ability to produce surprising outputs is a cornerstone of creativity, which in turn is widely considered to be an essential component of intelligent behavior (Boden, 1991). Surprise is a powerful driver for creativity and discovery. As such, surprise has been used to guide search algorithms in models of computational creativity and discovery (Yannakakis and Liapis, 2016). Owing to its importance for the creative process, surprise has also become one of the core criteria for the evaluation of creative artifacts (Maher et al., 2013). As reviewed in (Itti and Baldi, 2009), surprise is an essential concept in many studies on the neural basis of behavior, with surprising stimuli shown to be strong attractors of attention. Surprise, or violation of expectation, has also been hypothesized to be an essential mechanism through which music and stories elicit emotion. According to (Meyer, 1961), the principal emotional content of music arises from the composer's manipulation of expectation. Composers build expectations in time, which then they purposely violate in order to elicit tension, prediction, reaction, and appraisal responses (Huron, 2008). In text and narratives, surprise can be employed with substantial emotional impact at multiple levels, spanning from word-level, as in "Elon Musk has just blasted the world 's most powerful rocket into landfill" where the original word "space" was purposely replaced with "landfill" for humorous effect, to story-level, as in the various types of plot twists that are used to draw the reader emotionally in the story, e.g. *peripeteia* or *deux ex machina*.

In this section, we attempt to characterize word-level surprise using probability distributions computed by language models. We first consider a number of measures of surprise in the context of a general probability distribution p over an event space X, followed by more specialized surprise measures that are targeted to the special case of X being a language vocabulary. As such, we are interested in measuring how surprising the occurrence of an event $x \in X$ is for the audience p. An event x is improbable if its probability $p(x)$ is very small. Since improbable events are rare, it is tempting to consider the occurrence of an improbable event as being surprising. Weaver (1948) pondered on whether low probability implies surprise, *"an improbable event is often interesting. But is an improbable event always interesting?"*, and concluded *"we shall see that it is not"*, providing a simple, prototypical example in which improbable events are intuitively not surprising: a uniform distribution over an event space that has a large cardinality, as in dealing off a single bridge hand of thirteen cards from a shuffled pack of cards. There are more than 635 billion configurations of thirteen cards, all equally likely. Whatever bridge hand is dealt, although its probability is very small, it will not be, or feel, surprising. *"Any hand that occurs is simply one out of a number of exactly equally likely events, some one of which was bound to happen"*. What makes an event interesting or surprising is not that its probability is small in an absolute sense, it is that it is small in comparison to the probabilities of the other alternative events.

Weaver's insight is also in agreement with the interpretation of *"surprise as violation of expectation"*, which is hypothesized to be a major factor underlying emotion in music (Meyer, 1961). In this context, the term *expectation* refers to the kind that is engineered by composers in their music or by writers in their stories. Informally, a strong expectation is created when one or more potential outcomes are much more likely than other outcomes. More formally, an expectation regarding a random variable x is created when, prior to its value being observed, its context h makes a potential outcome $x = j$ more likely than other outcomes, as measured through the probability $p(x = j|h)$. Upon observing outcome $x = k$, we call it surprising if it confounds the expectation of seeing outcome $x = j$, i.e. $p(x = k|h) \ll p(x = j|h)$. Like in Weaver's argument above, the relative likelihood requirement for creating expectations immediately rules out uniform distributions.

The intuitive lack of surprise when observing

events sampled from a uniform distribution makes Shannon's *surprisal* inadequate as a measure of surprise. It is thus important that the notion of *surprise* is not equated with *surprisal*. The surprisal of an event x is an information-theoretic quantity defined as the negative log probability of x, i.e. $-\log p(x)$. Since surprisal is based solely on the event probability, monotonically decreasing with it, using surprisal to model surprise has the same conceptual deficiency as saying that rare events are surprising, as originally observed by Weaver (1948). Henceforth, to avoid confusion, we will refer to $-\log p(x)$ as the *information content* of x.

2.1 Quantifying Surprise

In this section, we describe a number of measures of surprise that are meant to capture the notion of small relative probability associated with surprising events. These measures are summarized in Table 1.

One of the first measures of surprise was the *surprise index* λ_1 proposed by Weaver (1948):

$$\lambda_1(p,x) = \frac{E[p]}{p(x)} \quad (1)$$

Weaver's surprise index is multiplicative: if X and Y are independent with distributions p and q, then the surprise index of the joint event $[x,y]$ is $\lambda_1(pq,[x,y]) = \lambda_1(p,x)\lambda_1(q,y)$.

Observing that the numerator $E[p]$ with which $p(x)$ is compared is somewhat arbitrary, Good (1956) generalized Weaver's surprise index to the following multiplicative (λ_c) and additive (Λ_c) versions, for $c > 0$:

$$\lambda_c(p,x) = \frac{(E[p^c])^{1/c}}{p(x)} \quad (2)$$

$$\Lambda_c(p,x) = \log \lambda_c(p,x) \quad (3)$$

Of all possible values for c, Good recommended as the most natural λ_0 and λ_1, together with their logarithmic versions Λ_0 and Λ_1, respectively:

$$\lambda_1(p,x) = \frac{E[p]}{p(x)} \quad (4)$$

$$\Lambda_1(p,x) = \log E[p] - \log p(x) \quad (5)$$

$$\lambda_0(p,x) = \frac{exp(E[\log p])}{p(x)} \quad (6)$$

$$\Lambda_0(p,x) = E[\log p] - \log p(x) \quad (7)$$

The additive measure Λ_0 is appealing because it can be interpreted in information theoretic terms as the difference between the Shannon information content $I(p,x) = -\log p(x)$ and the Shannon entropy $H(p)$:

$$\Lambda_0(p,x) = -\log p(x) - E[-\log p] \quad (8)$$

$$= I(p,x) - H(p) \quad (9)$$

Howard (2009) observes that Weaver's index can be written as $\lambda_1(p,x) = E\left[\frac{p}{p(x)}\right]$, whereas Good's index can be written as the mean of the log of the same variable, i.e. $\Lambda_0(p,x) = E\left[\log \frac{p}{p(x)}\right]$.

Observing that additive surprise indexes like Λ_0 more easily exceed a given value when the dimensionality is increased, Good (1988) advocated for using the tail-area probability as a surprise measure:

$$t(p,x) = \sum_{x':p(x')\leq p(x)} p(x') \quad (10)$$

However, the tail-area does not necessarily select outcomes that occur with small *relative probability*, for example when there are n alternative outcomes with slightly different probabilities that are all close to $1/n$. Howard (2009) points out that this behavior is connected to the fact that tail-area is not continuous in the outcome probabilities $p(x)$ and proposes a new measure of surprise called *s-value*:

$$sv(p,x) = 1 - \sum_{x'} min(p(x'),p(x)) \quad (11)$$

$$= 1 - [t(p,x) + n_x \cdot p(x)] \quad (12)$$

where n_x is the number of discrete outcomes with probability greater than $p(x)$. The s-value is continuous in $p(x)$ and, unlike the tail-area, selects for outcomes that conform with the basic intuition of small relative probability. It is equivalent with the probability mass contained in the area under the *pdf* curve that is above the $p(x)$ level.

If we use the term expectation with its psychological meaning of anticipation of an occurrence that may take place in future, a number of alternative definitions of surprise quantify the gap between the *psychological expectation* of a future event, i.e. the probability of the most likely event m_p, and its *realization*, i.e. the probability of the actual event x that happened. Correspondingly, the Expectation Realization (ER) gap can be defined as:

$$\psi(p,x) = \textit{Expectation} - \textit{Realization}$$

$$= \max_{x'} p(x') - p(x)$$

$$= p(m_p) - p(x) \quad (13)$$

Name	Formula	Unit
Good's surprise index	$\Lambda_0(p,x) = -\log p(x) - H(p)$	Information (bits)
Howard's s-value	$sv(p,x) = 1 - \sum_{x'} min(p(x'), p(x))$	Probability mass
Mode ER gap	$\Psi_m(p,x) = -\log p(x) + \log p(m_p)$	Information (bits)
Core ER gap	$\Psi_C(p,x) = -\log p(x) + \log p(C_p)$	Information (bits)

Table 1: Selected measures of surprise that capture the notion of small relative probability.

where $m_p = \arg\max_x p(x)$ is the largest mode of the distribution p, i.e. the expected, most likely outcome. Similar to Weaver and Good's surprise indexes, one can define a *multiplicative* version:

$$\begin{aligned} \psi(p,x) &= \textit{Expectation}/\textit{Realization} \\ &= p(m_p)/p(x) \end{aligned} \quad (14)$$

as well as an *additive* version:

$$\begin{aligned} \Psi_m(p,x) &= \log \textit{Expectation} - \log \textit{Realization} \\ &= \log p(m_p) - \log p(x) \\ &= I(p,x) - I(p,m_p) \end{aligned} \quad (15)$$

The ER measures for surprise are continuous in $p(x)$ and conform to the basic intuition of a surprising event having a small relative probability. We note that the simple ER gap from 14 has been previously proposed by Macedo et al. (2004), who found it to correlate well with human ratings of surprise. We prefer the additive version from 15 due to its information theoretic interpretation.

The measures of surprise proposed so far are summarized in the top 3 rows of Table 1. The measures were selected based on their properties, as follows: Good's surprise index and the Mode ER gap for their information-theoretic interpretation, and Howard's s-value for its probability mass interpretation. Of the 3 measures, the s-value and the mode ER gap also have the desirable property that they are non-negative for any outcome x, and become zero when x is the most likely outcome.

2.1.1 The Core Expectation Realization Gap

In this paper, we estimate surprise using the probability distribution computed by a language model. However, this creates a mismatch between the lexical level used to support the distribution and the semantic level that was used to annotate the creative examples. Most often, creativity implies surprise in terms of meaning, not necessarily in terms of the particular words chosen to express that meaning. Thus, the use of lexical distributions to estimate semantic surprise can lead to poor estimates of surprise in cases where a strong semantic expectation can be expressed with a large number of words. For example, to determine that "Congressmen" is surprising in the metaphor "an infestation of [Congressmen]", it is not sufficient that the realization $x =$ "Congressmen" in the context "an infestation of x" has a low probability. We also need a measure that tells us there is a strong expectation for what x is anticipated to be in the phrase "an infestation of x". In this example, the expectation is especially strong in terms of the semantic category of x, i.e. the reader strongly expects to see an instance from the PESTS category. Because this is a large category, there is a large set of words that can be reasonably expected in this context, resulting in a weak word-level expectation. Hence, the mode of the distribution used by the ER gap Ψ_m will not have a sufficiently high probability to make the Ψ_m pass a surprise threshold. The partition of the category expectation into many small word-level expectations leads to an increase in entropy, which adversely affects Good's surprise index Λ_0 as well.

For lack of an effective LM-based approach to compute probability distributions over semantic spaces, we designed an alternative version of the ER gap measure called Core ER gap, where the largest mode of the distribution m_p is replaced with the *Core* of the distribution C_p, comprising all the events $x \in X$ whose probability passes a predefined threshold, i.e. $C_p = \{x \in X | p(x) > \tau\}$. By appropriately setting the lower bound τ, we expect to capture in the core C_p all words belonging to the most expected semantic categories in a given context. Due to its information theoretic interpretation, we consider only the *additive* version:

$$\begin{aligned} \Psi_C(p,x) &= \log \textit{Expectation} - \log \textit{Realization} \\ &= \log p(C_p) - \log p(x) \\ &= I(p,x) - I(p,C_p) \end{aligned} \quad (16)$$

This version of the new Core ER gap measure is listed at the bottom of Table 1.

3 Datasets of Creative Language

We built two datasets of creative language examples: a HUMOR dataset and a METAPHOR dataset. The humor examples were extracted from the Humicroedit dataset (Hossain et al., 2019), which consists of regular English news headlines paired with versions of the same headlines that contain simple replacement edits designed to make them funny. Each funny headline was scored by five judges, resulting in a curated dataset of over 15,000 headline pairs. As positive examples for humor, we randomly selected 400 examples from a subset of the humorous headlines that were originally created using single-word replacements and that had an average annotator score of 1.8 or higher. The positive examples for metaphor were extracted from the English section of the LCC Metaphor dataset (Mohler et al., 2016) where the average annotator rating was 3.0 or above and where the source field of the metaphor was a single word. Furthermore, as explained below, we further applied a filtering step designed to preserve only metaphors that are novel to the language model, leaving a total of 268 positive examples of metaphor.

While a metaphor may appear creative to a person hearing it for the first time, it will sound completely unoriginal to a listener who has heard it and used it so many times that it has become part of their normal use of language. Similarly, a line that triggered laughter upon its first utterance, when repeated multiple times will normally get a smile at best from an audience already habituated to it. Therefore, it is important that creativity be determined with reference to a listener's experience. In general, judgements of creativity require specifying a reference model, e.g. the listener, the reader, or the audience, consuming the output produced by the speaker, the writer, or the composer, respectively. Consequently, based on the premise that creativity requires novelty, building an evaluation dataset annotated with creative uses of language requires fixing a *reference reader* and ensuring that examples annotated as creative are 1) novel for this reader and 2) evaluated with respect to the same reader. Since the proposed measures of surprise will necessitate access to the reader's contextual word distributions, in this paper we set the reference reader to be a generic reader whose knowledge of language is modeled by a large language model (LM), such as BERT (Devlin et al., 2019) if both the left and the right context of a word are used, or OPT (Zhang et al., 2022) if only using the previous discourse as context. Given that BERT was trained on the BooksCorpus and English Wikipedia, it is safe to assume that its pre-training data was not contaminated with any of the humorous headlines from Humicroedit, and therefore the humorous headlines appear novel to the reader modeled by BERT. However, we cannot say the same for the metaphor examples, as many of them are commonly used and likely to be found in BERT's pre-training corpus, e.g. "floating ideas", "deep understanding", "stealing dreams", "crushing insurgencies", "leap of faith", "seeds of discontent", to list just a few. To ensure that the metaphor examples included in the dataset are novel with respect to the reader modeled by BERT, since we did not have access to the exact pre-training data, we devised a conservative filtering where the base metaphor phrases were filtered out if a Google search returned less than 25 documents containing the phrase or its variations. For example, given the annotated metaphor "the bureaucracy barrier", we removed the article and also searched for "bureaucratic barrier" and "barrier of bureaucracy". Furthermore, we removed examples where the source word is repeated in the sentence context, as in "this [prison]$_s$ is the prison of [poverty]$_t$".

In terms of negative examples, for humor we used the 400 original titles corresponding to the 400 humorous examples. We further augmented these negative examples with nouns (as tagged by NLTK's POS tagger) selected at random from news articles downloaded from the CNN website in July 2022, such that the number of positive examples represents 10% of the total number of examples in each dataset. Regular news articles are expected to use regular language, without novel humor or novel metaphors. This is not to say the news articles do not contain metaphors, but when that happens they are metaphors that are commonly used and thus unsurprising for a generic reader. To summarize, the label distribution in the two datasets is as follows:

1. The Humor dataset, 4000 examples:
   (a) 400 positive examples, one-word substitution in news headlines that made them humorous, extracted from examples in the Humicroedit dataset with high inter-annotator agreement.
   (b) 400 negative examples, using the substituted word from the original titles used in the 400 positive examples above.

(c) 3200 negative examples, using random content words from CNN news articles.

2. The Metaphor dataset, 3760 eamples:

   (a) 268 positive examples, the annotated one-word source domain field of metaphors from the LCC Metaphor dataset that had high inter-annotator agreement and were rare on the internet.

   (b) 2412 negative examples, a subset of the 3200 selected at 1.(c) above.

The imbalanced label distribution was meant to address the fact that instances of creative language are relatively rare, although the exact proportion in general is hard to estimate due to the fact that certain types of text, e.g. poetry, are expected to be substantially more creative than other types, e.g. news articles. We note that the labels in the resulting dataset are likely to be noisy: metaphors that we annotated as creative, even though uncommon ad litteram on the internet, may still have been present in the LM's pre-training data in a different form, such as using a synonym for any of the words in the expression. Furthermore, it is possible that the CNN news articles included in the dataset contain instances of creative language, albeit very few. Overall though, it is expected that a good measure of surprise would show substantial discriminative power between the soft positive vs. soft negative examples in this dataset. Hardening the dataset would require the development of feasible annotation guidelines for determining whether the reference LM (the reference reader) has been exposed (through its pre-training data) to any given expression, and then going over each example and using the annotation criteria to determine the label.

4 Experimental Evaluation

All the distribution-based measures of surprise evaluated in this section were calculated using the probability distributions computed by the BERT Large model (cased) available on the HuggingFace website[3]. This is done by taking the word that is labeled in the dataset, masking it, and asking BERT to output the token distribution at the masked position, using a context size of 15 tokens to the left and to the right. Due to the WordPiece subword tokenization used by BERT, sometimes the word that need to be labeled is split into multiple tokens, where the first token is distinguished from the continuation tokens using the double hashtags '##', as for example 'disrespect' = 'di' + '##s' + '##res' + '##pect'. In these cases, we use the probability of the first token as a proxy for the probability of the entire word – preliminary experiments where the simple product or the geometric mean of all the token probabilities were used did not show a significant difference in the results, likely due to the fact that continuation tokens often receive a very high probability.

A starting assumption in these experiments is that the input text is well formed, e.g. it does not contain ungrammatical phrases or typos. While we recognize that real text may contain ill formed language that could be incorrectly detected as surprising by the various surprise measures proposed in this paper, we do not consider this to pose a significant challenge as such text could be feasibly detected and filtered out using current state-of-the-art NLP tools. Furthermore, a simple way to filter out ill formed language and typos is to ignore tokens that belong to the tail of the LM distribution, a procedure that we will investigate in future work.

The support of the raw LM distribution is modified to exclude *continuation tokens*, *non-content words*, and *punctuation symbols* for the reasons explained below, after which the probabilities of the remaining tokens are renormalized so that their total probability mass is still 1. Continuation tokens sometimes receive a high probability at the masked position. For example, in the annotated metaphor "$[\text{tax}]_t$ $[\text{sorcery}]_s$ is a mystery to me", when the source word "sorcery" is masked the continuation token "##ation" receives the highest probability, corresponding to the reasonable completion "taxation is a mistery to me". Since the masked word cannot be continuation in our task, all continuation tokens are eliminated from the distribution support. Depending on the context, non-content words such as determiners and prepositions may receive a high probability at the masked position, as for example in the metaphor text "we had our own little electoral "irregularities" down here in Portsmouth's First Ward, where we suffer from $[\text{constipated}]_s$ $[\text{democracy}]_t$". Determiners such as 'a' or 'the' receive a relatively high probability for occurring at the masked position for the source field. Since metaphors and one-word humorous word substitutions are content words, we remove non-content words from the distribution support. Punctuation symbols may also receive a relatively

[3] https://huggingface.co/bert-large-cased

Measures		creative Humor					creative Metaphor				
		P	R	F_1	F_{1m}	AuC	P	R	F_1	F_{1m}	AuC
Random baseline		10.0	50.0	16.7	–	–	10.0	50.0	16.7	–	–
All positive baseline		10.0	100.0	18.2	–	–	10.0	100.0	18.2	–	–
Information content	$I(p,x)$	32.0	86.5	46.7	50.3	46.2	27.6	79.2	40.8	47.8	38.2
Good's surprise index	$\Lambda_0(p,x)$	28.2	73.5	40.7	45.2	39.4	27.8	75.6	40.5	47.3	33.8
Howard's s-value	$sv(p,x)$	22.5	85.3	35.5	43.9	38.1	21.9	85.9	34.8	47.3	34.3
Mode ER gap	$\Psi_m(p,x)$	30.7	82.5	44.7	48.6	44.1	27.8	78.7	41.0	47.6	35.9
Core ER gap	$\Psi_C(p,x)$	31.6	85.8	46.2	49.8	45.6	27.8	79.2	41.1	48.1	38.3
Info $\wedge$ Entropy	$[I(p,x), H(p)]$	32.7	87.8	47.6	53.4	47.4	27.4	80.8	40.8	47.7	37.2
Info $\wedge$ Mode Info	$[I(p,x), I(p,m_p)]$	31.7	86.3	46.4	52.7	46.4	27.6	80.0	40.9	47.8	37.9
Info $\wedge$ Core Info	$[I(p,x), I(p,C_p)]$	33.0	88.3	48.0	53.2	49.3	27.7	79.5	41.0	48.1	38.2
Info $\wedge$ Entropy $\wedge$ Mode Info $\wedge$ Core Info		33.4	88.0	48.4	53.6	49.5	29.8	82.7	43.6	53.1	42.3
Contextual Embeddings + 2-layer FCN		80.3	89.8	84.5	87.1	91.2	93.7	94.1	93.7	95.1	95.6

Table 2: Results from comparative evaluation of surprise measures on detecting creative use of language.

high probability in some contexts, as such they are excluded as well from the distribution support. In the metaphor example "communism thrives on an empty stomach and [democracy]$_t$ [relaxes]$_s$ on a full one", symbols such as commas ',' and the dashes '–' are predicted with a high probability at the masked source position.

4.1 Quantifying Discriminative Power

To estimate the discriminative power of the various surprise measures, we use them as input features for a simple binary binary logistic regression model. During training of this linear classifier, given the imbalanced label distribution, positive examples are given 9 times the weight of negative examples in the cross-entropy cost function. Evaluation is done in a 10-fold setting, where each dataset is shuffled and partitioned into 10 equally-sized folds, then 9 folds are used as training and the remaining fold as testing. This training-testing procedure is repeated 10 times so that test results are obtained for each fold. Care was taken to ensure that test folds are not contaminated with information from training. Thus, metaphor examples that had the same target word were always placed in the same fold. The original title and the humorous title obtained by one-word substitution were also always placed in the same fold. Precision (P), recall (R), and F_1-measure are computed by pooling results across the 10 folds. Furthermore, by varying a threshold over the probabilistic output of the classifier, we create precision vs. recall graphs and use them to calculate two additional scores: the maximum F_1 measure across all confidence thresholds (F_{1m}) and the area under the curve (AuC).

4.2 Results and Discussion

For each dataset, Table 2 show the performance of 2 simple baselines, 5 standalone distribution-based measures, and 4 combinations of information-based measures. The 'random' baseline assigns labels uniformly at random, whereas the 'all positive' baseline labels every example as positive. In terms of combinations, for each of the 3 information measures we used the two terms in the measure as separate features. Therefore, since Good's surprise index is written as information content minus entropy, we evaluated a binary classifier that uses information content and entropy as two separate features. Similarly, the information content and mode information combination corresponds to the Mode ER Gap, whereas the information content and core information combination corresponds to the Core ER Gap. Finally, we use all these information terms as features in an overall combination, as shown at the bottom of the table.

The results show that all standalone measures do much better than random, showing that they do capture an important signal in terms of creative use of language. Somewhat surprisingly, no surprise measure does better than information content, despite the proven theoretical deficiency of using information content to model surprise. Of the 4 surprise measures, the Core ER Gap performs the best, being slightly under information content on Humor and slightly better than information content

on Metaphor. We hypothesize that an important reason for the lower performance of standalone surprise measures is the fact that the LM probabilities are miscalibrated. While calibration of probability distributions for classification tasks downstream of LM has been investigated in a number of recent works (Wang et al., 2020; Desai and Durrett, 2020; Park and Caragea, 2022), we are not aware of any work targeting calibration of the LM distribution itself. It is known for example that the tail of the LM distribution is unreliable (Holtzman et al., 2019), giving too much probability mass to words that should not be acceptable in the given context, e.g. resulting in ungrammatical phrases. The Mode and Core ER gaps ignore the the tail of the distribution completely, which may explain their relatively better performance when compared with Good's surprise index and Howard's s-value.

Since theoretically the average, mode, and core information are important for quantifying the level of surprise, instead of adding them directly to information content as was done in the surprise measures, we aimed to alleviate the miscalibration issue by training a linear model to optimize the trade-off between each of them and information content. The results in Table 2 show that, overall, when all types of information-based measure are combined, there is a substantial 3% increase in overall performance (AuC) over information content alone, on both datasets. The improvements in F_1 measure are statistically significant at $p < 0.01$, as measured using a one-tailed paired T-test over the results from the 10 folds. Overall, these results empirically support the theoretical observation that surprise measures capture aspects of creative language use that go beyond simple information content.

Finally, although the focus of this paper is on the discriminative power of surprise measures that are based solely on word-level distributions, the last line of Table 2 shows the performance of a classifier that uses the contextual representations produced by the frozen LM as input to a fully connected network (FCN) consisting of 2 hidden layers and one output logistic regression node. Unsurprisingly, the use of contextual embeddings as input to the FCN leads to much better results, likely due to its better capacity for modeling semantic-level surprise.

Humor $\vee$ ***Metaphor*** $\not\Longrightarrow$ ***Creative*** We would like to emphasize here that the detection of creative language evaluated in this section, although using examples drawn from humor and metaphor datasets, is quite different from the metaphor or humor detection tasks pursued in related work. The metaphor detection task (Leong et al., 2020) is unconcerned with whether the metaphor is commonly used vs. novel or surprising to the reader. In comparison, as argued in Section 3, creative language detection requires specifying a *reference reader* and the examples that are annotated as creative, be they humor or metaphor, need to be novel to this reader.

4.3 Error Analysis

Upon looking at the errors in which the trained classifier had the most confidence, we discovered a few major sources of errors. First, in terms of false negative, sometimes the metaphor word that is tagged as the source is made highly predictable by the presence of other words in the context, as in "[democracy]$_t$ is the thinly gloved [hand]$_s$ of repressive power", where the likelihood of hand is high due to the preceding 'gloved'. A possible solution could be to mask the entire phrase 'thinly gloved hand' when asking the LM to compute the probability distribution, and utilizing an encode-decoder LM such as T5 to produce a probability distribution over phrases. There also also instances of parallel metaphors in the same sentence, where one metaphor is highly predictive of the other, as in ""If [poverty]$_t$ is a [fire]$_s$ and aid is a firefighter, good governance is the water".

In terms of false positive, there are words that are associated with high information content because BERT does not have knowledge of named entities or types of events mentioned in the text. For example, in the title "Texas church [shooter] was Atheist, thought Christians stupid", the word shooter had a very low probability, likely due to BERT not having been trained on text referencing shootings in places of worship. Likewise, 'Harvey' receives a very low probability in the sentence "Trump has pledged $1 million to [Harvey] relief". In a way, these examples, although they were considered as negative by default, they are indeed surprising for the reference reader modeled by BERT.

5 Related Work

Owing to its essential role in our daily lives, there have been numerous computational approaches to humor recognition, as reviewed for example in (West and Horvitz, 2019; Hossain et al., 2019). Humor generation has presented a challenging problem in AI since the early 1990s, leading to the

development of various template-based and neural approaches (Amin and Burghardt, 2020). The important role that surprise plays in humor generation has been previously recognized in theories of humor, such as the surprise theory of laughter (Toplyn, 2014) and other prominent models that posit humor is evoked by incongruity within a text, such as the two-stage model of Suls (1972). According to incongruity theories of humor, a text conveys at least two interpretations, of which one is more salient. As readers process the text, the salient interpretation is activated until a text segment is encountered that contradicts it and thus promotes the previously unexpected interpretation. Surprise arises from his sudden revision of understanding.

Metaphors are pervasive in everyday communication, as well as in creative writing such as novels and poetry. Metaphors enhance the communicative aspects of language by connecting concepts from new domains, often abstract, with more familiar ones, usually concrete (Lakoff and Johnson, 1980). Metaphorical expressions have many uses, from helping frame an issue in order to emphasize some aspects of reality (Boeynaems et al., 2017), to creating a strong emotional effect (Blanchette and Dunbar, 2001; Citron and Goldberg, 2014). The ubiquity of metaphors means their computational treatment (Veale et al., 2016) has received significant attention in the NLP community, as surveyed by Shutova (2015) and more recently Tong et al. (2021). A distinction is made in the literature between *conventional* metaphors, which are entrenched in the conceptual system, and *novel* metaphors, which are unfamiliar. In this paper, we further recommend that novelty judgements be made relative to a *reference reader*. Our use of a large LM to model the reference reader is supported by the fact that pre-trained LMs encode conventional metaphorical information, as shown recently in the probing study of Aghazadeh et al. (2022). Even though metaphor is widely seen as a creative tool and surprise is an essential component of creative artifacts, we are not aware of any work investigating the role of surprise in discriminating between conventional vs. novel metaphors.

Computational approaches to humor and metaphor are part of a larger inquiry into identifying and formalizing the basic processes underlying human creativity. In the growing field of computational creativity[4], surprise has been proposed as one of the major criteria for the evaluation of creative artifacts (Maher et al., 2013). Surprising outputs were shown to attract the attention of the observer (Itti and Baldi, 2006), but also to guide the creative process itself: in a study of the creative design process followed by architects (Suwa et al., 2000), surprising discoveries in design sketches were observed to cause reformulations of design goals, which in turn led to further unexpected discoveries, due to designers reading more off a sketch than what they originally intended to put there (Schon and Wiggins, 1992). In this paper we emphasize that surprise, and by extension creativity, needs to be defined relative to a reference reader or audience. Consequently, generative architectures that aim to learn patterns of surprise and expectation from data need to contain a separate model for the reference reader, as implemented in the composer-audience models from (Bunescu and Uduehi, 2019) for binary sequences and (Uduehi and Bunescu, 2021) for basic geometrical shapes.

6 Conclusion and Future Work

Aiming to characterize creative language, we introduced a number of measures of surprise that are based solely on the probability distributions computed by a reference LM, considered to model a reference reader. Experimental evaluations show that, in combination with information content, the surprise measures improve detection of novel metaphors or humor, providing empirical evidence for the role of surprise in creative use of language. The code and data will be made publicly available[5].

Future work includes refining the datasets, calibrating the LM probabilities, developing semantic-level measures of surprise, and evaluating the proposed measures with respect to a reference reader that only knows the literal meaning of words. An interesting future extension to other types of word-level humor such as puns was suggested by a reviewer, where surprise measures would be combined with measures of character-level similarity such as edit distance.

Acknowledgements

We would like to thank the anonymous reviewers for their suggestions and constructive feedback.

[4] https://computationalcreativity.net

[5] https://github.com/uoseremen/SurpriseCreativeLanguage

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56. PMID: 15298329.

Isabelle Blanchette and Kevin Dunbar. 2001. Analogy use in naturalistic settings: The influence of audience, emotion, and goals. *Memory & Cognition*, 29(5):730–735.

Margaret A. Boden. 1991. *The Creative Mind: Myths and Mechanisms*. Basic Books, Inc., New York, NY, USA.

Amber Boeynaems, Christian Burgers, Elly Konijn, and Gerard Steen. 2017. The impact of conventional and novel metaphors in news on issue viewpoint. *International Journal of Communication*, 11(0).

Razvan Bunescu and Oseremen Uduehi. 2019. Learning to surprise: A composer-audience architecture. In *ICCC*, pages 41–48.

Francesca M. M. Citron and Adele E. Goldberg. 2014. Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of Cognitive Neuroscience*, 26(11):2585–2595.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

August Fenk and Gertraud Fenk-Oczlon. 1980. Konstanz im kurzzeitgedächtnis - konstanz im sprachlichen informationsfluß? *Zeitschrift für experimentelle und angewandte Psychologie*, 27:400–414.

Futrell R. Piantadosi S. P. Dautriche I. Mahowald K. Bergen L. Levy R. Gibson, E. 2019. How efficiency shapes human language. *Trends in cognitive sciences*, 23(5):389—407.

I. J. Good. 1956. The Surprise Index for the Multivariate Normal Distribution. *The Annals of Mathematical Statistics*, 27(4):1130 – 1135.

I. J. Good. 1988. Surprise index. *Encyclopedia of Statistical Sciences*, 7(1):1–5.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The Curious Case of Neural Text Degeneration.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

J. V. Howard. 2009. Significance testing with no alternative hypothesis: A measure of surprise. *Erkenntnis (1975-)*, 70(2):253–270.

David Huron. 2008. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT.

Laurent Itti and Pierre Baldi. 2009. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295 – 1306.

Laurent Itti and Pierre F. Baldi. 2006. Bayesian surprise attracts human attention. In *NIPS*. MIT Press.

T. Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press, Chicago.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Luis Macedo, R. Reisezein, and A. Cardoso. 2004. Modeling forms of surprise in artificial agents: empirical and theoretical study of surprise functions. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*.

Mary Lou Maher, Katherine A. Brady, and Douglas H. Fisher. 2013. Computational models of surprise in evaluating creative design. In *Proceedings of the Sixth International Conference on Computational Creativity (ICCC)*.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Typical Decoding for Natural Language Generation. *arXiv:2202.00666 [cs]*. ArXiv: 2202.00666.

Leonard Meyer. 1961. *Emotion and Meaning in Music*. University of Chicago.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Seo Yeon Park and Cornelia Caragea. 2022. On the Calibration of Pre-trained Language Models using Mixup Guided by Area Under the Margin and Saliency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5364–5374, Dublin, Ireland. Association for Computational Linguistics.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Donald A Schon and Glenn Wiggins. 1992. Kinds of seeing and their functions in designing. *Design studies*, 13(2):135–156.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

C. E. Shannon. 1951. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30(1):50–64.

Ekaterina Shutova. 2015. Design and Evaluation of Metaphor Processing Systems. *Computational Linguistics*, 41(4):579–623. _eprint: https://direct.mit.edu/coli/article-pdf/41/4/579/1807226/coli_a_00233.pdf.

Jerry M. Suls. 1972. Chapter 4 - A Two-Stage Model for the Appreciation of Jokes and Cartoons: An Information-Processing Analysis. In JEFFREY H. GOLDSTEIN and PAUL E. McGHEE, editors, *The Psychology of Humor*, pages 81–100. Academic Press, San Diego.

Masaki Suwa, John Gero, and Terry Purcell. 2000. Unexpected discoveries and s-invention of design requirements: Important vehicles for a design process. *Design Studies*, 21(6):539–567.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. Association for Computational Linguistics.

Joe Toplyn. 2014. *Comedy Writing for Late-Night TV: How to Write Monologue Jokes, Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-Form Comedy*.

Oseremen O. Uduehi and Razvan C. Bunescu. 2021. Adversarial learning of expectation and surprise: Experiments with geometric shapes. In *ICCC*, pages 286–290.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A Computational Perspective. *Synthesis Lectures on Human Language Technologies*, 9(1):1–160. Publisher: Morgan & Claypool Publishers.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics.

Warren Weaver. 1948. Probability, rarity, interest, and surprise. *The Scientific Monthly*, 67(6):390–392.

Robert West and Eric Horvitz. 2019. Reverse-Engineering Satire, or "Paper on Computational Humor Accepted despite Making Serious Advances". *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7265–7272. Number: 01.

Georgios N. Yannakakis and Antonios Liapis. 2016. Searching for surprise. In *Proceedings of the Seventh International Conference on Computational Creativity*, pages 25–32, PAris, France.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.