How Creative is Your Writing? A Linguistic Creativity Measure from Computer Science and Cognitive Psychology Perspectives

Xiaojin Zhu, Zhiting Xu and Tushar Khot
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI, USA 53706
{jerryzhu, zhiting, tushar}@cs.wisc.edu

Abstract

We demonstrate that subjective creativity in sentence-writing can in part be predicted using computable quantities studied in Computer Science and Cognitive Psychology. We introduce a task in which a writer is asked to compose a sentence given a keyword. The sentence is then assigned a subjective creativity score by human judges. We build a linear regression model which, given the keyword and the sentence, predicts the creativity score. The model employs features on statistical language models from a large corpus, psychological word norms, and WordNet.

1 Introduction

One definition of *creativity* is "the ability to transcend traditional ideas, rules, patterns, relationships, or the like, and to create meaningful new ideas, forms, methods, interpretations, etc." Therefore, any computational measure of creativity needs to address two aspects simultaneously:

1. The item to be measured has to be different from other existing items. If one can model existing items with a statistical model, the new item should be an "outlier".

2. The item has to be meaningful. An item consists of random noise might well be an outlier, but it is not of interest.

In this paper, we consider the task of *measuring human creativity in composing a single sentence, when the sentence is constrained by a given keyword*. This

simple task is a first step towards automatically measuring creativity in more complex natural language text. To further simplify the task, we will focus on the first aspect of creativity, i.e., quantifying how *novel* the sentence is. The second aspect, how *meaningful* the sentence is, requires the full power of Natural Language Processing, and is beyond the scope of this initial work. This, of course, raises the concern that we may regard a nonsense sentence as highly creative. This is a valid concern. However, in many applications where a creativity measure is needed, the input sentences are indeed well-formed. In such applications, our approach will be useful. We will leave this issue to future work. The present paper uses a data set (see the next section) in which all sentences are well-formed.

A major difficulty in studying creativity is the lack of an objective definition of creativity. Because creative writing is highly subjective ("I don't know what is creativity, but I recognize it when I see one"), we circumvent this problem by using human judgment as the ground truth. Our experiment procedure is the following. First, we give a keyword z to a human writer, and ask her to compose a sentence $\mathbf{x}$ about z. Then, the sentence $\mathbf{x}$ is evaluated by a group of human judges who assign it a subjective "creativity score" y. Finally, given a dataset consisting of many such keyword-sentence-score triples $(z, \mathbf{x}, y)$, we develop a statistical predictor $f(\mathbf{x}, z)$ that predicts the score y from the sentence $\mathbf{x}$ and keyword z.

There has been some prior attempts on characterizing creativity from a computational perspective, for examples see (Ritchie, 2001; Ritchie, 2007;

Pease et al., 2001). The present work distinguishes itself in the use of a statistical machine learning framework, the design of candidate features, and its empirical study.

2 The Creativity Data Set

We select 105 keywords from the English version of the Leuven norms dataset (De Deyne and Storms, 2008b; De Deyne and Storms, 2008a). This ensures that each keyword has their norms feature defined, see Section 3.2. These are common English words.

The keywords are randomly distributed to 21 writers, each writer receives 5 keywords. Each writer composes one sentence per keyword. These 5 keywords are further randomly split into two groups:

1. The first group consists of 1 keyword. The writers are instructed to "write a not-so-creative sentence" about the keyword. Two examples are given: "Iguana has legs" for "Iguana", and "Anvil can get rusty" for "Anvil." The purpose of this group is to establish a non-creative baseline for the writers, so that they have a sense what does not count as creative.

2. The second group consists of 4 keywords. The writers are instructed to "try to write a creative sentence" about each keyword. They are also told to write a sentence no matter what, even if they cannot come up with a creative one. No example is given to avoid biasing their creative thinking.

In the next stage, all sentences are given to four human judges, who are native English speakers. The judges are not the writers nor the authors of this paper. The order of the sentences are randomized. The judges see the sentences and their corresponding keywords, but not the identity of the writers, nor which group the keywords are in. The judges work independently. For each keyword-sentence pair, each judge assigns a subjective creativity score between 0 and 10, with 0 being not creative at all (the judges are given the Iguana and Anvil examples for this), and 10 the most creative. The judges are encouraged to use the full scale when scoring. There is statistically significant ($p < 10^{-8}$) linear correlation among the four judges' scores, showing their general agreement on subjective creativity. Table 1 lists the pairwise linear correlation coefficient between all four judges.

Table 1: The pairwise linear correlation coefficient between four judges' creativity scores given to the 105 sentences. All correlations are statistically significant with $p < 10^{-8}$.

	judge 2	judge 3	judge 4
judge 1	0.68	0.61	0.74
judge 2		0.55	0.74
judge 3			0.61

The scores from four judges on each sentence are then averaged to produce a consensus score y. Table 2 shows the top and bottom three sentences as sorted by y.

As yet another sanity check, note that the judges have no information which sentences are from group 1 (where the writers are instructed to be non-creative), and which are from group 2. We would expect that if both the writers and the judges share some common notion of creativity, the mean score of group 1 should be smaller than the mean score of group 2. Figure 1 shows that this is indeed the case, with the mean score of group 1 being 1.5 ± 0.6, and that of group 2 being 5.1 ± 0.4 (95% confidence interval). A t-test shows that this difference is significant ($p < 10^{-11}$).
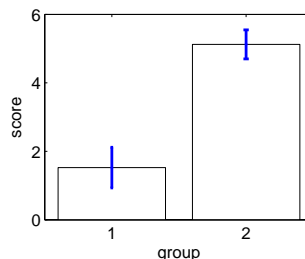

Figure 1: The mean creativity score for group 1 is significantly smaller than that for group 2. That is, the judges feel that sentences in group 2 are more creative.

To summarize, in the end our dataset consists of 105 keyword, sentence, creativity score tuples $\{(z_i, \mathbf{x}_i, y_i)\}$ for $i = 1, \ldots, 105$. The sentence group information is not included. This "Wisconsin Creative Writing" dataset is publicly available at `http:`

Table 2: Example sentences with the largest and smallest consensus creativity scores.

consensus score y	keyword z	sentence $\mathbf{x}$
9.25	hamster	She asked if I had any pets, so I told her I once did until I discovered that I liked taste of hamster.
9.0	wasp	The wasp is a dinosaur in the ant world.
8.5	dove	Dove can still bring war by the information it carries.
...		
0.25	guitar	A Guitar has strings.
0.25	leech	Leech lives in the water.
0.25	elephant	Elephant is a mammal.

//pages.cs.wisc.edu/~jerryzhu/pub/
WisconsinCreativeWriting.txt.

3 Candidate Features for Predicting Creativity

In this section, we discuss two families of candidate features we use in a statistical model to predict the creativity of a sentence. One family comes from a Computer Science perspective, using large-corpus statistics (how people *write*). The other family comes from a Cognitive Psychology perspective, specifically the word norms data and WordNet (how people *think*).

3.1 The Computer Science Perspective: Language Modeling

We start from the following hypothesis: if the words in the sentence $\mathbf{x}$ frequently co-occur with the keyword z, then $\mathbf{x}$ is probably not creative. This is of course an over-simplification, as many creative sentences are about novel usage of common words[1]. Nonetheless, this hypothesis inspires some candidate features that can be computed from a large corpus.

In this study, we use the Google Web 1T 5-gram Corpus (Brants et al., 2007). This corpus was generated from about 10^{12} word tokens from Web pages. It consists of counts of N-gram for $N = 1, \ldots, 5$. We denote the words in a sentence by $\mathbf{x} = x_1, \ldots, x_n$, where $x_1 = \langle s \rangle$ and $x_n = \langle /s \rangle$ are special start- and end-of-sentence symbols. We

[1]For example, one might argue that Lincoln's famous sentence on government: "of the people, by the people, for the people" is creative, even though the keyword "government" frequently co-occurs with all the words in that sentence.

design the following candidate features:

[f_1: **Zero N-gram Fraction**] Let $c(x_i^{i+N-1})$ be the count of the N-gram $x_i \ldots x_{i+N-1}$ in the corpus. Let $\delta(A)$ be the indicator function with value 1 if the predicate A is true, and 0 otherwise. A "Zero N-gram Fraction" feature is the fraction of zero N-gram counts in the sentence:

$$f_{1,N}(\mathbf{x}) = \frac{\sum_{i=1}^{n-N+1} \delta(c(x_i^{i+N-1}) = 0)}{n - N + 1}. \quad (1)$$

This provided us with 5 features, namely N-gram zero count fractions for each value of N. These features are a crude measure of how surprising the sentence $\mathbf{x}$ is. A feature value of 1 indicates that none of the N-grams in the sentence appeared in the Google corpus, a rather surprising situation.

[f_2: **Per-Word Sentence Probability**] This feature is the per-word log likelihood of the sentence, to normalize for sentence length:

$$f_2(\mathbf{x}) = \frac{1}{n} \log p(\mathbf{x}). \quad (2)$$

We use a 5-gram language model to estimate $p(\mathbf{x})$, with "naive Jelinek-Mercer" smoothing. As in Jelinek-Mercer smoothing (Jelinek and Mercer, 1980), it is a linear interpolation of N-gram language models for $N = 1 \ldots 5$. Let the Maximum Likelihood (ML) estimate of a N-gram language model be

$$p_{ML}^N(x_i|x_{i-N+1}^{i-1}) = \frac{c(x_{i-N+1}^i)}{c(x_{i-N+1}^{i-1})}, \quad (3)$$

which is the familiar frequency estimate of probability. The denominator is the count of the history of length $N - 1$, and the numerator is the count of the history plus the word to be predicted. A 5-gram

Jelinek-Mercer smoothing language model on sentence $\mathbf{x}$ is

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|x_{i-5+1}^{i-1}) \qquad (4)$$

$$p(x_i|x_{i-5+1}^{i-1}) = \sum_{N=1}^{5} \lambda_N P_{ML}^N(x_i|x_{i-N+1}^{i-1}), (5)$$

where the linear interpolation weights $\lambda_1 + \ldots + \lambda_5 = 1$. The optimal values of λ's are a function of history counts (binned into "buckets") $c(x_{i-N+1}^{i-1})$, and should be optimized with convex optimization from corpus. However, because our corpus is large, and because we do not require precise language modeling, we instead set the λ's in a heuristic manner. Starting from N=5 to 1, λ_N is set to zero until the N where we have enough history count for reliable estimate. Specifically, we require $c(x_{i-N+1}^{i-1}) > 1000$. The first N that this happens receives $\lambda_N = 0.9$. The next lower order model receives 0.9 fraction of the remaining weight, i.e., $\lambda_{N-1} = 0.9 \times (1 - 0.9)$, and so on. Finally, λ_1 receives all remaining weight to ensure $\lambda_1 + \ldots + \lambda_5 = 1$. This heuristic captures the essence of Jelinek-Mercer smoothing and is highly efficient, at the price of suboptimal interpolation weights.

[f_3: **Per-Word Context Probability**] The previous feature f_2 ignores the fact that our sentence $\mathbf{x}$ is composed around a given keyword z. Given that the writer was prompted with the keyword z, we are interested in the novelty of the sentence surrounding the keyword. Let x_k be the first occurrence of z in the sentence, and let $\mathbf{x}_{-k}$ be the *context* of the keyword, i.e., the sentence with the k-th word (the keyword) removed. This notion of context novelty can be captured by

$$p(\mathbf{x}_{-k}|x_k = z) = \frac{p(\mathbf{x}_{-k}, x_k = z)}{p(x_k = z)} = \frac{p(\mathbf{x})}{p(z)}, \quad (6)$$

where $p(\mathbf{x})$ is estimated from the naive Jelinek-Mercer 5-gram language model above, and $p(z)$ is estimated from a unigram language model. Our third feature is the length-normalized log likelihood of the context:

$$f_3(\mathbf{x}, z) = \frac{1}{n-1}\left(\log p(\mathbf{x}) - \log p(z)\right). \qquad (7)$$

3.2 The Cognitive Psychology Perspective: Word Norms and WordNet

A text corpus like the one above captures how people *write* sentences related to a keyword. However, this can be different from how people *think about related concepts* in their head for the same keyword. In fact, common sense knowledge is often underrepresented in a corpus – for example, why bother repeating "A duck has a long neck" over and over again? However, this lack of co-occurrence does not necessarily make the duck sentence creative.

The way people think about concepts can in part be captured by *word norms* experiments in psychology. In such experiments, a human subject is provided with a keyword z, and is asked to write down the first (or a few) word x that comes to mind. When aggregated over multiple subjects on the same keyword, the experiment provides an estimate of the concept transition probability $p(x|z)$. Given enough keywords, one can construct a concept network where the nodes are the keywords, and the edges describe the transitions (Steyvers and Tenenbaum, 2005). For our purpose, we posit that a sentence $\mathbf{x}$ may not be creative with respect to a keyword z, if many words in $\mathbf{x}$ can be readily retrieved as the norms of keyword z. In a sense, the writer was thinking the obvious.

[f_4: **Word Norms Fraction**] We use the Leuven dataset, which consists of norms for 1,424 keywords (De Deyne and Storms, 2008b; De Deyne and Storms, 2008a). The original Leuven dataset is in Dutch, we use a version that is translated into English. For each sentence $\mathbf{x}$, we first exclude the keyword z from the sentence. We also remove punctuations, and map all words to lower case. We further remove all stopwords using the Snowball stopword list (Porter, 2001), and stem all words in the sentence and the norm word list using NLTK (Loper and Bird, 2002). We then count the number of words x_i that appear in the norm list of the keyword z in the Leuven data. Let this count be $c_{norm}(\mathbf{x}, z)$. The feature is the fraction of such norm words in the original sentence:

$$f_4(\mathbf{x}, z) = \frac{c_{norm}(\mathbf{x}, z)}{n}. \qquad (8)$$

It is worth noting that the Leuven dataset is relatively small, with less than two thousand keywords. This

is a common issue with psychology norms datasets, as massive number of human subjects are difficult to obtain. To scale our method up to handle large vocabulary in the future, one possible method is to automatically infer the norms of novel keywords using corpus statistics (e.g., distributional similarity).

[$f_5 - f_{13}$: **WordNet Similarity**] WordNet is another linguistic resource motivated by cognitive psychology. For each sentence **x**, we compute Word-Net 3.0 similarity between the keyword z and each word x_i in the sentence. Specifically, we use the "path similarity" provided by NLTK (Loper and Bird, 2002). Path similarity returns a score denoting how similar two word senses are, based on the shortest path that connects the senses in the hypernym/hyponym taxonomy. The score is in the range 0 to 1, except in those cases where a path cannot be found, in which case -1 is returned. A score of 1 represents identity, i.e., comparing a sense with itself. Let the similarities be $s_1 \ldots s_n$. We experiment with the following features: The mean, median, and variance of similarities:

$$f_5(\mathbf{x}, z) = \text{mean}(s_1 \ldots s_n) \qquad (9)$$
$$f_6(\mathbf{x}, z) = \text{median}(s_1 \ldots s_n) \qquad (10)$$
$$f_7(\mathbf{x}, z) = \text{var}(s_1 \ldots s_n). \qquad (11)$$

Features $f_8, \ldots, f_{12}$ are the top five similarities. When the length of the sentence is shorter than five, we fill the remaining features with -1. Finally, feature f_{13} is the fraction of positive similarity:

$$f_{13}(\mathbf{x}, z) = \frac{\sum_{i=1}^{n} \delta(s_i > 0)}{n}. \qquad (12)$$

4 Regression Analysis on Creativity

With the candidate features introduced in Section 3, we construct a linear regression model to predict the creativity scores given a sentence and its keyword.

The first question one asks in regression analysis is whether the features have a (linear) correlation with the creativity score y. We compute the correlation coefficient

$$\rho_i = \frac{\text{Cov}(f_i, y)}{\sigma_{f_i} \sigma_y} \qquad (13)$$

for each candidate feature f_i separately on the first row in Table 3. Some observations:

- The feature f_4 (Word Norms Fraction) has the largest correlation coefficient -0.48 in terms of magnitude. That is, the more words in the sentence that are also in the norms of the keyword, the less creative the sentence is.

- The feature f_{12} (the 5-th WordNet similarity in the sentence to the keyword) has a large positive coefficient 0.47. This is rather unexpected. A closer inspection reveals that f_{12} equals -1 for about half of the sentences, and is around 0.05 for the other half. Furthermore, the second half has on average higher creativity scores. Although we hypothesized earlier that more similar words means lower creativity, this (together with the positive ρ for f_{10}, f_{11}) suggests the other way around: more similar words are correlated with higher creativity.

- The feature f_5 (mean WordNet similarity) has a negative correlation with creativity. This feature is related to f_{12}, but in a different direction. We speculate that this feature measures the strength of similar words, while f_{12} indirectly measures the number of similar words.

- The feature f_3 (Per-Word Context Probability) has a negative correlation with creativity. The more predictable the sentence around the keyword using a language model, the lower the creativity.

Next, we build a linear regression model to predict creativity. We use stepwise regression, which is a technique for feature selection by iteratively including / excluding candidate features from the regression model based on statistical significance tests (Draper and Smith, 1998). The result is a linear regression model with a small number of salient features. For the creativity dataset, the features (and their regression coefficients) included by stepwise regression are shown on the second row in Table 3. The corresponding linear regression model is

$$\hat{y}(\mathbf{x}, z) = -5.06 \times f_4 + 1.80 \times f_{12} - 0.76 \times f_3$$
$$-3.39 \times f_5 + 0.92. \qquad (14)$$

A plot comparing $\hat{y}$ and y is given in Figure 2. The root mean squared error (RMSE) of this model is

Table 3: ρ: The linear correlation coefficients between a candidate feature and the creativity score y. β: The selected features and their regression coefficients in stepwise linear regression.

	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$	$f_{1,5}$	f_2	f_3	f_4	f_5
ρ	0.09	0.09	0.17	0.06	-0.04	-0.07	-0.32	-0.48	-0.41
β							-0.76	-5.06	-3.39

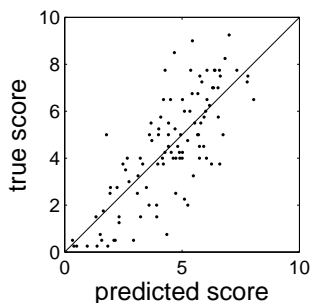|	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}
ρ	-0.19	-0.25	-0.02	0.06	0.23	0.30	0.47	-0.01
β							1.80	

Figure 2: The creativity score $\hat{y}$ as predicted by the linear regression model in equation 14, compared to the true score y. Each dot is a sentence.

1.51. In contrast, the constant predictor would have RMSE 2.37 (i.e., the standard deviation of y).

We make two comments:

1. It is interesting to note that our intuitive features are able to partially predict subjective creativity scores. On the other hand, we certainly do not claim that our features or model solved this difficult problem.

2. All three kinds of knowledge: corpus statistics (f_3), word norms (f_4), and WordNet (f_5, f_{12}) are included in the regression model. Coincidentally, these features have the largest correlation coefficients with the creativity score. The fact that they are all included suggests that these are not redundant features, and each captures some aspect of creativity.

5   Conclusions and Future Work

We presented a simplified creativity prediction task, and showed that features derived from statistical language modeling, word norms, and WordNet can partially predict human judges' subjective creativity scores.

Our problem setting is artificial, in that the creativity of the sentences are judged with respect to their respective keywords, which are assumed to be known beforehand. This allows us to design features centered around the keywords. We hope our analysis can be extended to the setting where the only input is the sentence, without the keyword. This can potentially be achieved by performing keyword extraction on the sentence first, and apply our analysis on the extracted keyword.

As discussed in the introduction, our analysis is susceptible to nonsense input sentences, which could be predicted as highly creative. Combining our analysis with a "sensibility analysis" is an important future direction.

Finally, our model might be adapted to explain why a sentence is deemed creative, by analyzing the contribution of individual features in the model.

6   Acknowledgments

References

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *EMNLP*.

S. De Deyne and G Storms. 2008a. Word associations: Network and semantic properties. *Behavior Research Methods*, 40:213–231.

S. De Deyne and G Storms. 2008b. Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, 40:198–205.

Norman R. Draper and Harry Smith. 1998. *Applied Regression Analysis (Wiley Series in Probability and Statistics)*. John Wiley & Sons Inc, third edition.

Frederick Jelinek and Robert L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Workshop on Pattern Recognition in Practice*.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *The ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 62–69.

Alison Pease, Daniel Winterstein, and Simon Colton. 2001. Evaluating machine creativity. In *Workshop on Creative Systems, 4th International Conference on Case Based Reasoning*, pages 129–137.

Martin F. Porter. 2001. Snowball: A language for stemming algorithms. Published online.

Graeme Ritchie. 2001. Assessing creativity. In *Proceedings of the AISB01 Symposium on Artificial Intelligence and Creativity in Arts and Science*, pages 3–11.

Graeme Ritchie. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17(1):67–99.

Mark Steyvers and Joshua Tenenbaum. 2005. The large scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.