Conflict Classification via a Probe based on a Pretrained LLM
A Subtask of Humor Detection
ANLP23 Final Project Report

Kai Hu, Bofei Xu
UC Berkeley

1 Introduction

You hear something funny, you laugh. Although, you won't sit there, judge the material, and go "Oh it has the necessary components A, B, and C. Therefore, it qualifies as a funny joke. I should laugh. Hahaha!!". There must exist certain linguistic structures or characteristics in the material that stimulated our brain, such that our synapses would fire up, and our laughter would burst out. What are these structures and components? Numerous researchers have proposed linguistic models that attempt to answer this question. How do machines determine if a text is funny? Previous researchers have put forth many humor detection machines, such as the transformer based classifier by Weller and Seppi (2019), and the multimodal contexual extension of Memory Fusion Network (C-MFN) by Hasan et al. (2019). One of the ability of Neural Network based models is to extract implicit features in the form of the activation vectors in the hidden layers. While the classifiers are able to distinguish humorous texts from non-humorous texts, do they adhere to a certain linguistic model that, researchers believe, human brains do? This paper aims to answer part of that question.

The first step of this study was to establish how human brains works, for which we resorted to humor linguistic theories. Within the framework of the reviewed humor linguistic theories, we specified the probing task as a sub-task of humor detection, which we refer to as conflict-type classification. We will discuss this in the following sections in more details.

We trained the probe using the GridLoc method (Niu et al. (2022)), which outputs weights assigned to each token. We performed statistic testings, including the Mann-Whitney U test, and the Bi-serial test on the weights to test our hypothesis on the distributional differences between different group of tokens, as well as the correlation between the weights and the binary importance as labeled by the annotators.

2 Related Work

In this section, we presented the related works by previous researchers. The literature review focused on both the linguistic humor theories, as well as the probing techniques.

2.1 Linguistic Humor Model - Incongruity Resolution

The incongruity-resolution theory of humor is widely advocated for, and is the basis of many variants. Due to the lack of standard terminologies, and the inherent difficulty in the definition of incongruity, many variants of the IR model exist, out of which two of the more clearly defined (Ritchie (1999)) are the surprise disambiguation model (as summarized by Ritchie of notions contained in Shultz (1974), Minsky (1984), and Paulos (1980)), and the two-stage model (Suls (1972)).

2.1.1 Surprise Disambiguation

For the SD model, Ritchie proposed that a humorous text include three entities M1, "the first (more obvious) interpretation of the set-up", M2, "the second (hidden) interpretation of the set-up", and M3 "the meaning of the punchline", as well as five inter-relations/properties of these entities, obviousness, "M1 is more likely than M2 to be noticed by the reader", conflict, "M3 does not make sense with M1", compatibility, "M3 does make sense with M2", comparison, "there is some contrasted relationship, even a clash between M1 and M2." , and inappropriateness, "M2 is inherently odd, eccentric, or preposterous, or is taboo."

2.1.2 Suls' Two-Stage Model

As summarized by Ritchie (1999), the two-stage model process the text with the follow procedure: The reader predict the next words given the read words. When no conflict is encountered with the reader's prediction, the reader keep going. When a conflict is encountered, if the text is not ended, the reader will be puzzled. If the text is ended, then further, if the reader can find a cognitive rule that resolves the conflict then humor is found, otherwise the reader is also left in puzzlement.

2.2 Probing

Probing is an approach used to study the internal representations of Neural Networks. With a pretrained neural-network based humor classifier (or a classifier trained for a even more generous purpose), or in another word a pretained encoder, we plan to use probing to to measure whether the characteristic linguistic structures of humorous texts are captured in the implicit features extracted by the encoder. Tenny et al. (Tenny et al. (2019)) employed the "edge probing" method combined with eight established probing tasks (Conneau et al. (2018)). They used te scalar mixing technique (Peters et al. (2018)) to pool the activation vectors, or under NLP setting the contextualized token embeddings, at the internal hidden layers of the encoder, as well as the original non-contextualized token embedding vectors. Then the resulted "weighted average" of these vectors (or the pooled vectors) are used to make prediction by the probing classifiers. The original encoding are not changed, which means when training the probing classifiers only the scalar parameters including the weights for each hidden layers, and a overall scaling parameter, were learned, not the parameters of the pretrained encoder. This is to ensure that the neural network won't have just simply learned the probing task by modification of its encoding, but, again, have already included this information in its encoding.

This probing architecture was used by Tenny et al. (2019), not only to perform the probing tasks, but more importantly to show in which layers are the probing tasks solved, in the claim that BERT rediscovered the classical NLP pipeline (upstream layers accountable for surface tasks, then layers at intermediate depths for syntactic tasks, finally downstream layers for semantic tasks). However, in an reappraisal (Niu et al. (2022)) of this effort, the authors showed that "pseudo-cognitive appeals to layer depth may not be the preferable mode of explanation for BERT's inner workings", by showing that the layer depths do not have strong statistical correlation to the type of probing tasks. Given the controversy and our different aim than Tenny et al. (2019), we do not emphasize in which layers are the information stored but rather focus on the existence of it. Niu et al. (2022) used an alternative "GridLoc" probing architecture, which not only assign attention weights to the hidden layers, but also dynamically assign attention weights to tokens. This allowed the analyses of the token importance in solving a probing task. This is significant to us for reason that will become clear in later sections. Therefore we plan to perform experiments using the GridLoc probing technique.

3 Formalization of the Probing Task

Within the surprise disambiguation model, there are two types of conflicts. The first type is the conflict between the punchline and the first more obvious interpretation of the set up. The second type is the conflict between the punchline and other more natural/expected predictions made by the viewer given the set-up.

An example of a joke containing the first type of conflict:

- "Why do birds fly south in winter?
  It's too far to walk." Ritchie (1999)

The second type of conflict can be observed in the following joke.

- "Isn't modern technology wonderful? I remember the excitement when we were the first family in our street to have cordless pyjamas. (Arnold Brown, early 1990s)" Ritchie (1999)

In this study, we annotated the humorous texts by the type of conflicts they contain, and had this binary classification task as the probing task.

This is a sentence level task, which we used the aforementioned GridLoc probing technique to investigate, on a token level, which tokens attracted the most attention across the layers, when performing this text sequence level task. The GridLoc method enabled us to not only investigate whether this knowledge was present in the pretained NN, but also which tokens were the most important in making that classification.

4 Dataset

There is not a readily available annotated dataset suitable for training the proposed probe. Therefore, we annotated the Short Jokes dataset (Moudgl, 2017), which contains 231,657 jokes, with varying length from 10 to 200 characters. This dataset was built by scraping the subreddits /r/jokes and /r/cleanjokes. The jokes were downloaded from the day of posting to 31st Jan, 2017. We reviewed the jokes with IDs from 1 to 1,487, and from 231,313 to 231,657, which is a total of 1,832 jokes.

We labeled the jokes by the types of conflict they contain as described in the earlier sections. The labels entail:

- Type 0 - A joke, of which the punchline conflicts with the first more obvious interpretation of the set-up.

- Type 1 - A joke, of which the punchline conflict with the viewer's prediction of the punchline given the setup.

We determined that 189 jokes out of the reviewed jokes contained the two types of conflicts (The annotators found the others are in general not very funny anyways). 80 belongs to Type 0; 109 belongs to Type 1. The class ratio between the conflict types is 0.73 (Type 0 /Type 1).

The annotated dataset was processed into a text file with three columns per the input requirement by the GridLoc script, with the partitions ("tr" for training, "va" for validation, "te" for testing) in the first column, ground-truth classes in the second column, and the short jokes in the third column. The ratio between training, validation, and testing is 0.7, 0.15, and 0.15.

Additionally, during the annotation, we also labeled the important sequences. More concretely, for type 0 jokes, we labeled the set-up with two different interpretations, and the punchline; for type 1 jokes, we labeled the punchline. This is not for training a second probe, but for the statistic testings we will discuss in more detail in later sections.

5 Method

We employed a stock BERT model (Devlin et al. (2019)), "bert-base-uncased", as the pre-trained large language model. BERT, which is short for Bidirectional Encoder Representations from Transformers, was originally trained to predict the masked token given the rest of the contextual tokens in the text sequence. The underlying assumption is that by having been trained on a large set of corpora, BERT extracted syntactic, and semantic features as its internal text representation, which we can harvest with the architecture of the probes to perform downstream tasks.

We employed the GridLoc method in training the conflict-type probe classifier.

During training, the parameters are initialized randomly. Due to the effect of randomness in the trained probe parameters, which in turn determines the weights, we performed training with 100 random seeds and performed statistic testings on the weights. We pooled the trained weights obtained with all random seeds for the statistical testings.

With the labeling additional to the joke classes during annotation, we first grouped the tokens into the following six categories.

- Group 1 - The rest of the tokens in type 1 jokes, all other tokens excluding punctuation

- Group 2 - Punchline tokens of type 1 jokes, the tokens within the punchline of type 1 jokes

- Group 3 - The rest of the tokens in type 0 jokes, all other tokens excluding punctuation

- Group 4 - Punchline tokens of type 0 jokes, the tokens within the punchline of type 0 jokes

- Group 5 - Set-up tokens of type 0 jokes, the tokens within the set-up with two interpretations in type 0 jokes

To further illustrate the distinctions between group 1, 2, 3, 4, and 5, for example, in the type 0 joke, "Two guys walk into a bar. The third guy ducks.", tokens from "into a bar" are under group 5, the token of "ducks" is under group 4, and the ones from "Two guys walk", and "The third guy" are lumped into group 3; in the type 1 joke, "He was a real gentlemen and always opened the fridge door for me.", tokens of "fridge door" are under group 2, while the ones from "He was a real gentlemen and always opened the ", and "for me" are lumped under Group 1.

We aimed to find the distributional difference between these five token groups, for which we employed the Mann-Whitney U test (Mann and Whitney (1947)). The Mann-Whitney U test is a nonparametric test with the null hypothesis being that by randomly generating a value X from distribution A, and a value Y from distribution B, the probability of X being greater than Y, is equal to the probability of the opposite.

We also assigned binary importance to the tokens. We considered tokens in group 2, 4, and 5 important in classifying the conflict type the joke contains, which we gave an importance of 1. Tokens from Group 5 are considered relatively not as important, and were given an importance of 0. In order to study the correlation between the weights and the binary importance as labeled by the annotators, we employed the Point-Biserial test. The Point-Biserial test, which is a special case of Pearson's R test, was used to obtain a correlation coefficient which measures the correlation between a dichotomous variable X, and a continuous variable Y. The coefficient was taken to prove or reject the null hypothesis of that there is not a significant correlation between X and Y.

6 Analysis

6.1 Mann-Whitney U Test

Mann-Whitney U Test is used to study the distributional difference between the token weights of each groups. Five categories of tokens are collected through the following procedures:

- Gather attention weight matrices for 28 testing sentences using 100 different seeds during training initialization.
- Consolidate the attention weights from the last layer for each testing sentence and store them in a data frame (the data frame's dimensions should be $100\times$ the number of tokens in a sentence).
- Classify tokens in a sentence into five categories based on the index of important sequence labeling.
- Store five categories of attention weights into five arrays and ready for statistical testing.

Based on the Mann-Whitney U statistical test, the comparison results across all groups are found to be statistically significant. This provides strong evidence to reject the null hypothesis, suggesting that there is indeed a difference between the two tested groups. Moreover, differences within each type of joke are even more pronounced.

Specifically, the test results indicate a low p-value between groups 1 and 2 (both from Type 1 jokes). This suggests that pre-trained Language Models (LLMs) exhibit varying levels of attention based on the importance of tokens in a joke. The significance of this difference underscores the model's ability to discriminate between token importance.

Groups	P-value	Significant
1 vs 2	8.702895×10^{-77}	True
1 vs 3	2.925829×10^{-10}	True
1 vs 4	5.451308×10^{-4}	True
1 vs 5	8.422516×10^{-3}	True
2 vs 3	3.965515×10^{-57}	True
2 vs 4	5.524734×10^{-42}	True
2 vs 5	2.906200×10^{-23}	True
3 vs 4	4.800479×10^{-18}	True
3 vs 5	3.496586×10^{-2}	True
4 vs 5	1.059500×10^{-5}	True

Table 1: Mann-Whitney U Statistical Comparison

However, the test result between groups 3 and 5(both from Type 0 jokes)shows a higher p-value, approaching the 0.05 threshold for rejection. This may be attributed to the unique characteristics of the set-up category and the non-punchline category. Notably, most set-up tokens are either surrounded by or could be perceived as part of non-punchline tokens. This intricacy in token association may contribute to the less decisive test outcome for group 5.

6.2 Point-Biserial Correlation Coefficient

The point-Biserial Correlation Coefficient is used to study the strength and direction of the association between important token weights and regular token weights.

- Important tokens: Group 2, 4, and 5 tokens
- regular tokens: Group 1 an 3 tokens
- Array 1: Containing all token weights from all groups
- Array 2: Binary array where elements from important tokens are encoded in 1 and the rest as 0.

Point-Biserial Correlation Coefficient is measured between Array 1 and 2, with a coefficient of -0.0759, and a P-value of 2.3310^{-85}. The test result and P-value suggest we have strong evidence to reject the null hypothesis that there is no correlation between two arrays. However, the coefficient is close to 0 which implies the correlation between the importance of a token and the attention weight is weak. A negative coefficient indicates a negative relationship between the importance of token and attention weight, a token with a higher value of attention weight tends to be a regular token.

6.3 Testing accuracy

The pre-trained LLM achieved a median testing accuracy of 0.828 with a standard deviation of 0.194. Considering the relatively small training sample size, the model performs well in test accuracy. However, the standard deviation is higher than the expectation due to the limitation of training size. In general, the pre-trained LLM is capable of classifying Type 0 and Type 1 jokes as expected. The internal representation from pre-trained LLM contains hidden semantic and synthetic features that can be used to classify different types of jokes. The statistical result from the probing task supports our assumption that hidden semantic and synthetic features are used to classify jokes.

7 Conclusion and Limitations

In this study, we performed literature review in investigation of existing linguistic humor models, and probing techniques, and established a new probing task of conflict-type classification.

To train the probe, we reviewed and labeled a portion of the *short joke* data set.

We adopted the GridLoc probing method, and trained a probe classifier based on the pre-trained LLM, BERT, more specifically "bert-base-uncased". The median test accuracy was 0.828, with a standard deviation of 0.194, considering the 100 randomized training instances.

Additionally, we performed Mann-Whitney U tests, and rejected all pair-wise null hypotheses, which means significant distributional differences were found between all token groups as expected.

We also performed point-biserial tests and found significant correlation between the weights and the binary importance of the tokens.

Some limitations of this study include:

- Human annotation is subject to subjectivity even with adequate annotating specifications. There are only two annotators, whose perspective and therefore conclusions on the conflict types can be highly biased. To reduce the bias, a bigger group of annotators may be employed.

- Although the annotators were able to review a total of 1,832 joke, only approximately 10 percent of the reviewed jokes were suitable for this probing task of interest. This could lead to over fitting on the training data.

- Due to time constraints, this study used a stock BERT model. Ideally, BERT should be trained on a suitable large data set to perform a humorous text detection task before its parameters are frozen for the probe training. This could potentially improve the performance of the probe classifier in the test and validation accuracy, as well as alter the distributional differences between the token groups and the level of correlation between the weights and binary importance.

References

Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

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 (NAACL)*, pages 4171–4186.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Henry B. Mann and Donald R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60.

Marvin Minsky. 1984. *Jokes and the Logic of the Cognitive Unconscious*, pages 175–200. Springer Netherlands, Dordrecht.

Abhinav Moudgl. 2017. Short jokes dataset. https://github.com/amoudgl/short-jokes-dataset. Accessed: [Nov 10, 2023].

Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. Does BERT rediscover a classical NLP pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

John Allen Paulos. 1980. *Mathematics and Humor*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Graeme Ritchie. 1999. Developing the incongruity-resolution theory. Technical report.

Thomas R. Shultz. 1974. Development of the appreciation of riddles. *Child Development*, 45(1):100–105.

Jerry M. Suls. 1972. Chapter 4 – a two-stage model for the appreciation of jokes and cartoons: An information-processing analysis.

Ian Tenny, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, arXiv:1905.05950v2. Version 2.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.