Computational Humor: A Survey of the Literature

Vineeth NC
University of Maryland, Baltimore County
Baltimore, MD 21250 USA
vineethnc@umbc.edu

Abstract

Humor, a quintessentially human experience, is one that finds itself at the intersection of psychology, linguistics, philosophy, and computation. Thus, it finds itself in a very unique place in the world of computer science, especially when it comes to learning it through computational linguistics. The research so far is deep and significant, with work spanning across understanding, quantifying, and generating humor. In this paper, a survey of relevant literature is presented and discussed with the hopes of identifying the path that lays ahead and the work that is to be done to further develop a theory of humor that encompasses its computational understanding as well.

1 Introduction

The complexity in computational humor stems right at the very basic level: the definition of humor. For as long as language has existed, humor has too. Discussions of humor by researchers in different fields has given rise to many different definitions, or non-definitions in some cases, of humor. One definition comes from relief theory, which implies humor is a mechanism to enable relief of the psychological tension through laughter (Morreall, 2020). The loose, and often varied, definition of humor makes it almost impossible to construct steadfast rules or guidelines to identifying something as humor. Another important aspect concerning humor is that it is heavily reliant on context. A statement, or a group of statements, that may be considered humorous can cease to be so when separated from the context that they are originally attached with. (Attardo, 2010), in Chapter 0, discusses a few definitions proposed by linguists and psychologists, and the importance of subdivision of the concept of humor into different types. An important point brought up in the chapter is the fact that subcategorizing humor is difficult because of the possible dimensions to explore based on who is doing the categorization. While some fields conflate different terms like ”humor” and ”comedy” and ”comic” under the same umbrella, other fields tend to separate them based on the subjects they tackle, and the cultural context in which the words are being used.

The difference between the perception of humor across different fields, therefore, calls for a more thorough and collaborative approach to researching it. (Lewis, 1989) calls for a more interdisciplinary approach to researching humor, urging literary critics and social scientists to work in tandem. With advances in technology, researchers from fields like psychology and neurology have been able to work at this intersection, thereby adding to our understanding of humor. In the current technological landscape, it would be prudent to consider Lewis' call and have computer scientists join this intersection.

Humor has a very wide vocabulary, from verbal to written to funny sounds to physical slapstick humor. However, for the purposes of this paper, we will be focusing on humor that is centered around using language, be it verbal or written. While a literature review does little in the way of proposing a novel idea, or a exciting new direction for the research to go in, it is still a significant undertaking. A literature review, especially for a topic that is this complex and intricate, collates important research and ideas across different fields and connects them, thereby bringing some order to the chaos that is the world of research. This, in turn, helps make it simpler for the research in the future to look up relevant information in a single place, with the opportunity to choose what topics or ideas to dig deeper into. This paper, on a high level, categorizes the research into three broad categories: humor detection and generation. The research papers are reviewed and common underlying themes are identified to help chart potential directions for the research to take.

2 Early Work

(Ritchie, 2001) notes that the discussions of humor ranges across a wide range of perspectives because humor as a topic has several different aspects or themes that all warrant research into. These themes are very diverse and unique in their nature, thereby spawning discourse that is just as varied. To express this, Ritchie cites a couple of ideas that are polar opposites of each other but are still true in their own regard. The first idea describes humor from a Freudian standpoint, where motivation for humor is said to be from a mental space that handles inappropriate thoughts and feelings. By extension, the jokes are said to inherently contain imperfect reasoning, with the actual literary structure of the joke playing second fiddle to its impact on the listener. The second idea contrasts the first, in that it considers the jokes' literary structure, but does not contain enough information to extrapolate a fully fleshed out image of the joke's social context. Inferences regarding different aspects of the context can be made, but the literary structure of the joke seldom in solidifying a hypothesis surrounding the context. Discourses like this only serve to show that no single theory can possibly encapsulate the subjective and interpretive nature of a witticism and, by extension, humor as a whole.

This very nature of the field of humor is what has been forcing computer scientists to take a highly refined, narrowed, and targeted approach to tackling it as a part of research. Attempts have been made to create metrics and ideas that are broadly applicable to the field as a whole. An interesting example of this is the attempt at quantification of ”sense of humor” (Suslov, 2007), where the formulation for a computer model is presented. This model attempts to combine the concept of information with the psychological processing of humor to meaningfully quantify the humorousness of a joke. While the scope of the model is quite limited, it is an interesting and a novel approach to start with.

This highlights a good starting point for the field of computational humor: how does a machine know what is funny, and what isn't? The identification of humor is a significant task for machines to be able to perform. To this end, (Ritchie, 1999) proposed a theory called the Incongruity-Resolution (henceforth IR) theory that attempts to generalize the general structure of a joke. The hope here is that this structure, when turned to a computational model, would help machines recognize humor with reasonable efficiency and accuracy. The theory is based off of (Beattie, 1776)'s essay, which first used the word incongruous to define the idea of humor as being a result of multiple incongruous parts, the peculiar relation between which catches the attention of the human mind and results in laughter. According to (Ritchie, 1999), the punchline is an attempt at resolving the incongruity of the joke by using information presented earlier in the joke, which then gives rise to humor. The processing of the joke contains multiple steps: the processor has to first analyze the setup of the joke. This then leads the processor to predict what are the most probable lexical tokens that can work as a continuation of the setup. After the prediction, the process must detect and process the punchline, and must end with understanding the humorous aspect of the punchline. As the author notes in their paper, it is not a trivial task to quantify and understand the humorous aspect of the joke, because that determines the resolution of the incongruity of the joke. An improperly constructed method could very well lose out on the nuance of the joke, and therefore misinterpret the punchline, and by extension the joke itself.

(Ritchie, 2001) mentions that as of the writing of the paper, the research and implementation in the field of computational humor was still in its infancy, with only a handful of interesting projects were successful in simulating humor mechanisms. The different projects mentioned in the paper range from concepts like analysis of irony to detecting puns to generating riddles based on puns. Since these projects are rudimentary at best, they are limited in their approach and implementation. However, there is a clear underlying theme connecting a lot of the research mentioned: a lot of the work is centered around humor that is verbally presented. These projects are all reliant on and demonstrate different levels of understanding of the IR theory, thereby lending more credibility to the theory itself. The IR theory, for the purposes of this paper, will be quite useful because it has motivated a lot of research since it was proposed as it helps set a theoretical framework for computational models to work with.

3 Humor Processing and Detection

(Mihalcea and Strapparava, 2005) explores the possibility of using computational approaches to recognize humor expressed in a verbal form. In their paper, the authors attempt to formulate the task of humor recognition as a binary classification task, where the positive label indicates that a given example is humorous, and the negative label indicates otherwise. The paper cites research that mentions certain features observed through the analysis of a significant number of jokes. It is observed that for shorter joke formats in a verbal medium, the sounds of the words have more impact on the audience. To this end, the authors recognized alliteration as a significant indicator of humor, and built an index that reflects the chains of alliterations present in the example.

The paper also extends the idea of incongruity in the construction of the humor, especially looking into finding conflicting statements where the punchline resolves the incongruity by going against the setup of the joke. This allows the model to look for words that can be considered to be antonyms. By finding such relationships between different lexical tokens in the setup and the punchline, the model can, to a certain extent, determine the humorous nature of the example in question. Another feature that they considered is the usage of adult language, which is a very popular format of humor all across the world, across most languages. Using all these features as heuristics to train Naive Bayes and Support Vector Machines classifiers, the authors were able to achieve reasonable efficiency and accuracy in their predictions on a test dataset of 15000 examples. One important takeaway from this is that this paper has demonstrated a way to consider specific features of a given example and be able to quantify them to help make an effective computational model.

The authors also conduct an experiment with different classifiers and training datasets to check the relation between training dataset sizes and classification model performance. The experiment showed that the classification performance plateaus after a certain training dataset size, implying that th ere may be an optimal training dataset size to build a computational model for recognizing humor and classifying examples. The approach to detecting humor as a classification problem motivated (Liu et al., 2018) to attempt to identify whether a given text contains expressions that can be considered humorous. The paper uses features defined in (Yang et al., 2015) to build a humor recognition system that is based on the idea of using sentiment analysis to recognize humor. This paper is an interesting extension to (Mihalcea and Strapparava, 2005) because it uses the same dataset as the latter, thereby making it easier to understand the difference in performance.

A key aspect of (Liu et al., 2018)'s work centers around the assumption that sentiment association would help in revealing the nature of humor in a given text example. A parser to extract discourse relations from a given body of text is used to help represent as a hierarchical structure called a Rhetorical Structure Theory (RST) type relation over the whole text. The leaf nodes, in this structure, are independent and individual text units that connect with other text units to form a discourse tree. The parser, after separating the sentences into different units, can establish relationships between each other, thereby creating a data structure on a corpus of text with relationships between different lexical units already identified. An important aspect of these features is the polarity of each individual unit in the tree, which helps the parser identify whether two units have opposite polarity. The existence of units with polar opposite sentiments is a good indicator of incongruity, which is used as an indicator of humor. The experiment described in the paper uses features mentioned earlier as baseline features and uses them in conjunction with sentiment associations to build a new classifier and compare its performance against the baseline classifier.

The results of the experiment are very interesting. It was observed that, in most cases, sentiment association improved the classifier performance across different metrics like accuracy and F1 score. Sentiment association seemingly manages to encode enough information about the relationships between different units in the text for the classifier to be able to interpret and understand the text better. The authors extended the classifier's training process by adding a new feature called emotional word count, which is a count of the words that have a positive or a negative polarity. When compared head-to-head, it is observed that sentiment association has more impact than emotional word counts in the performance of the classifier. While this is a novel idea, it might be a better approach to have an emotional word density feature as opposed to the emotional word count feature. Simply put, emotional word density is the ratio of the count of emotional words to the length of the text example. This is a good way to normalize the score because a higher emotional word count does not always indicate that the text itself contains more emotional content. A text with longer length and higher emotional word count could still have lesser density of emotional words when compared to a smaller text. It would be interesting to repeat this experiment with the ratio in conjunction with sentiment association to see if the density makes more of a difference than count, and if it boosts the performance or serves to be detrimental.

4 Humor Generation

What can computers do once they start understanding what is funny, and what isn't? The natural answer to that question is for them to be able to make jokes from their understanding of humor. (Ritchie, 2009) tackles this question by referring to the results from the paper cited in the previous section, (Mihalcea and Strapparava, 2005), to start questioning what makes humorous texts different from normal texts. He observes that the experiment conducted by (Mihalcea and Strapparava, 2005) demonstrates that there is no single specific determiner that distinguishes a humorous text from a non-humorous text, and that it might be as efficient to guess whether a text is humorous or not. While the experiment and its results are interesting, he notes that there is a long way to go before a humor creation system can be modeled and implemented. This leads us to the question that is the heart of this section: what does it take to generate humor? The answer to this is complex because modeling an algorithm has to consider a few things, and has to meet a certain standard in its output.

The author rightfully argues that a learning algorithm needs to be able to consider a corpus of text, learn from it, and generate output that is unique in its form and content while also being humorous. The reason he calls it a significant achievement is because the alternatives to such a model are much worse. Improper modeling can lead to a model that does not understand humor, or at best repackages the input data into texts that are shallow and mirror the input too closely for it to be unique. To illustrate this, the author talks about some of the first attempts at computational humor generators: pun generation. Puns are a very specific form of humor that involve manipulating words and syllables to generate humor. The narrow and specificity in the nature of puns leads the generator to have modules that perform specific tasks, like setting up the premise, generating the pun for the punchline. While this is a good example for how generation could work, the author notes that a general purpose humor generation model is one that can generate different kinds of humor with the same efficiency without being too fine-tuned to generate a single type of humor.

Currently, there is no single general purpose humor generating model, but many different researchers have come up with models that can generate more specific types of humor with a greater range of input. One such project is (Horvitz et al., 2020), an attempt at creating a model that can generate satirical headlines that incorporates the concept of context into humor generation. As noted earlier in this paper, context is a significant motivator of humor. Jokes taken out of context seldom retain the aspect that makes them witty, and either lose all meaning or become open to misinterpretation. The authors of this paper considered an existing news summarization architecture, BertSum, and finetune it based on a custom constructed dataset where the context is manually constructed by retrieving information from relevant real-world stories and events. This involves finding a way to model satirical news headlines with relation to their real-world context. The authors had to establish a pipeline involving human judges that could retrieve context for a given output. The exact details of how to elicit judgments about the texts from appropriate judges (for example, children for child-oriented humor) need to be planned, but this area is not mysterious, as this type of study is routine within psychology.

For this experiment, the authors used the popular satirical news website, The Onion, to get a dataset of headlines. These headlines were then fed through the information retrieval pipeline, which generated context for each of them. Once the context-centric dataset is generated, it is passed as the training dataset to the BertSum model, which upon training should be able to generate headlines of its own. However, understanding the performance of BertSum and the importance of context calls for an almost poetically fitting requirement of further context surrounding model performance. To achieve this, the authors chose to train a context-free headline generating model using the then popular GPT-2 architecture.

It is at this juncture of the paper that the question of model evaluation must be dealt with. Who/what decides whether a headline is funny or not? (Ritchie, 2009) touches upon the topic of testing of humor generating models. A model to evaluate a given input must have an understanding of humor that spans across different kinds of humor that uses human language. The computer models for testing, and the pipeline surrounding them must be free of human bias to consider any input given and objectively evaluate the humor presented. To build such a system is far in the future because a humor evaluation model is an inherent extension of the task of humor detection, followed by a quantification of the humor using a metric similar to (Suslov, 2007) but with more variety of inputs and understanding of humor. In the absence of such a system, (Ritchie, 2009) proposes that the best possible way for now is to use human judges. While humor is subjected to bias of the observer, evaluating humor is not too unlike surveys done in the field of psychology, which routinely deal with conducting experiments in a controlled setting with a specific group of testers that is large enough to render individual biases insignificant in the larger scheme of things. And this is the approach used by (Horvitz et al., 2020) for evaluating the context based satirical headlines generated by BertSum.

The results of human evaluation of the BertSum and GPT-2 outputs point to some interesting features of humor. The overarching takeaway from the experiment is that the evaluators found context-driven headlines to be funnier than context-free approaches, which supports the theoretical idea that humor is reliant around the context that it is a part of. Unsurprisingly, it was also observed that context-driven headlines included more ways of introducing incongruity, including ways that could be considered absurd and yet contextually relevant. It seemed that including context in the training process helped the model pick from a larger vocabulary of relevant lexical tokens in the construction of headlines, implying that it understood incongruity and humor better than the context-free model did. This model is a good example of the potential room for improvement that (Ritchie, 2009) mentions when he talks about the creativity without being radically original in the humor that is being generated.

5 Conclusion and Future Directions

Over the last forty years, a lot of progress has been made in understanding humor by people of different fields. It could be argued that (Ritchie, 1999)'s Incongruity-Resolution theory is one of the most seminal works in the field of computational humor, considering its influence and legacy across the field. Computational humor as a field is still young when compared to most other fields, but it is quickly gaining depth, thanks to the development of neural networks and artificial intelligence. In such a rapidly changing technological landscape, it is important that the assumptions being made stand the test of time and be proven right over and over again to ensure that the field is built on a strong foundation. Incremental work is being performed at a significant rate and old concerns are being addressed, as seen by (Horvitz et al., 2020)'s addressing of the issue of contextual relevance in humor generation raised by (Ritchie, 2009).

The future of computational humor is very bright because there is a long way to go. From the projects discussed in this paper, one potential direction that can be mapped out is to incorporate sentiment association alongside context for humor generation. Sentiment association might limit the vocabulary that a humor generation model can draw from, but can help make the token choices better by giving importance to tokens with higher association that heighten the incongruity and relieve it without losing relevance or meaning.

An important step towards creating a general-purpose humor model is to understand and quantify the relation between different types of humor. This can help a model detect and classify humor into one or more types of humor, and can also help in more efficient generation as well. While the computation will likely increase exponentially, it is a worthy undertaking because it can help shed more light into the intersection of fields that humor operates in, thereby prompting more research across different fields.

References

S. Attardo. 2010. *Linguistic Theories of Humor*. Humor Research [HR]. De Gruyter.

J. Beattie. 1776. *An Essay on Laughter and Ludicrous Composition, Written in the Year 1764*. W. Creech and E. and C. Dilly.

Zachary Horvitz, Nam Do, and Michael L. Littman. 2020. Context-driven satirical news generation. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 40–50, Online. Association for Computational Linguistics.

P. Lewis. 1989. *Comic Effects: Interdisciplinary Approaches to Humor in Literature*. State University of New York Press.

Lizhen Liu, Donghai Zhang, and Wei Song. 2018. Modeling sentiment association in discourse for humor recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 586–591, Melbourne, Australia. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

John Morreall. 2020. Philosophy of Humor. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2020 edition. Metaphysics Research Lab, Stanford University.

Graeme Ritchie. 2001. Current directions in computational humour. *Artificial Intelligence Review*, 16.

Graeme Ritchie. 2009. Can computers create humor? *AI Magazine*, 30:71–81.

Graeme D. Ritchie. 1999. Developing the incongruity-resolution theory. In *Proceedings of the AISB 99 Symposium on Creative Language: Humour and Stories*.

I. M. Suslov. 2007. Computer model of a ”sense of humour”. i. general algorithm. *ArXiv*, abs/0711.2058.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.