# 3. Connecting the Dots

Soham Roy - 7028704 -soro00002@stud.uni-saarland.de

## 1. Introduction and similarities in the problem space

*In this report, the keyword "first paper" refers to the paper [2] "Connecting the Dots Between News Articles", the "second paper" refers to the paper [3]"Inside Jokes: Identifying Humorous Cartoon Captions" and the "third paper" refers to the paper [1] "Accelerating Innovation Through Analogy Mining"*

The first paper explores methods for automatically connecting news articles to discover hidden connections between them. Finding a coherent chain linking them together, like a chain of events that has a causal effect on the next event so that readers can draw a connectivity to the events. The second paper focuses on the language of cartoon captions and its influence on humor perception. The third paper investigates the use of structural representations, specifically "problem schemas," to find useful analogies in large idea repositories.

All three papers discuss the use of computational methods to analyze and extract useful information from datasets that have been put together for some specific purposes and needed a lot of manual intervention and annotation to make the datasets for the problem specifics. They share a common paradigm in a way that the authors try to model human behavior and experience in a way by analyzing and finding patterns from datasets that have been obtained by crowdsourcing or public submissions. The author broke down the data into two sets depending on the problem at hand, such as context and anomaly in one case, and purpose and mechanism in the other. There are similarities in which both were analyzed. It was attempted to identify the parameters that make a joke funnier, such as perplexity and length, in the same joke context. In the second case, the author looked for two or more mechanisms that could achieve the same purpose (purpose as a context). In the third case, the author sought to determine the next document that would best fit the story or plot. All three problems attempt to find the best fit and match the existing data in hand to meet certain criteria based on some evaluation techniques. By keeping a factor constant, for Eg. the joke context or purpose, the author attempted to find the best suitable solution in the problem space from the other set which is the anomaly set or the mechanism set.

## 2. Methods and Assumptions

In paper two, the author created three clusters, each cluster denoting the same joke but in a different way, picked 10 jokes from each cluster and then did separate them by length, lengths of 5- 6 and 9 -10, length was chosen to bucket jokes as it was assumed that humor has a correlation with length! we do not get any info on participants like culture, age, ethnicity, financial status etc and many jokes tend to breed in these fields and can be funny to one group while not to another, also some jokes need prior knowledge to understand like typical stereotypical jokes, if one is not aware of that then he won't get the joke so it is important to address the jokes selection and their funniness ratings could be different based on crowds. Here it is probably assumed that it is a neutral crowd and a caption would be funnier to everyone in the universe.

When the author processes the tokens it is not mentioned how the punctuations are handled, punctuation can play a major role in the formation of a joke for Eg if using punctuation "?" eg I invited the clown, Trump to the party. Without the comma, it would imply that Trump is the clown.

The author studied various features and evaluated the importance of different features using the Gini importance score and found that pronouns, question words, negation words, and auxiliary verbs were the top discriminators, which was surprising given that many of these words are usually considered "stopwords."

The author replicates the image into two sets of tags called context words and anomalies, then computed the cosine similarity of the main joke phrase between the context and anomaly tags, trying to see whether the joke phrase is similar to context or anomaly, the visual image plays an important role which hasn't been addressed and has been mapped to a set of tags which is not the correct representation of how the elements (or tags) in the image interact with each other. The author selects each word of the joke phrase and uses it in the objective function to find where it lies on the anomaly-context scale, the author is ruling out the possibility that 2 or n-grams could be a potential check in the objective function.

The author doesn't address how perplexity is handled for two synonymous words which can differ a lot but do not

impart any new info in terms of funniness, so the causation shown between perplexity and funniness could be spurious. Here the author has calculated - **Funniness(simple word | joke = x)** and **Funniness (complex word |joke = x)** but it would be interesting to know how the other words in both the sentence played out with regard to the words in comparison (simple word and complex word), and since it is a sequence it is wrong to assume independency of the words to the other words .
Computing both lexical and syntactic distinctiveness can provide a more comprehensive analysis of what makes a joke funny, as humor often relies on both the choice of words and the structure of the sentence. The accuracy of the approach heavily depends on the accuracy of the POS tagging process, which may not always be perfect.

In paper one, the method is dependent on the chronology of reporting and has no ability to distinguish earlier occurring events being reported at a later date so completely ignores it, it can get only the partial truth and as the chain grows longer it begins to more and more starts to get out of context, for eg if the story is only 10 percent lined to the next one then longer chains tends to become vague as the length increases. It did a good job of looking far ahead in the line, unlike the shortest path approaches which ties only to the next document and tends to stray away from the objective. However, the method matches the exact words in the method and it is unclear about the words that are synonyms of each other but used in the different documents then it would fail to relate the words together unlike the other two papers which use words embedding which has the ability to capture semantic meaning. the author uses Copernic to find influence weights over the graph, Alternatively, the author has suggested using tf idf weights, this might not be a good idea cause it would weigh down the importance of important words depending on the number of occurrences on the document and can increase the importance of silly words, for eg. if document emphasizes on "Obama" and mention it many times while all of them gives important info, tf-idf will deem that less important and does not have the ability to consider that all of the occurrences has some new info associated with it, this will cause the "word" to "document" edge in the influence graph a high value to unimportant document and this in turn is the random Markov model transition probability.
Also, here the user has to manually enter source and sink nodes as it might be a challenge to come up with the right source and sink for a story without prior knowledge. As a constraint of the Linear Programming approach it has been mentioned that the Word activation level cannot exceed activation in the original document and since it is a continuous variable for a word if it is non-zero in its first document then the method is constraining the activation for this word in the successive documents. It would have been nice for the method to find parallel stories for eg - $A \rightarrow$B $\rightarrow$C and $A \rightarrow$D $\rightarrow$C, which could be $A \rightarrow$(B and D) $\rightarrow$C.
In paper three, it has been assumed that the pieces of text have one purpose and one mechanism. The method selects a list of keywords that categorizes it as purpose/ mechanism or neither. Suppose some text has more than one purpose let's say P1 and P2 and one mechanism M1, then the selection strategy has no way to distinguish that the keywords belong to P1 or P2 and it might end up selecting some from either (depending upon the value of D) which could lead to a not a poor purpose representation to either of them.

## 3. Results and Evaluation

In the first paper, Since solving the constraints by Linear programming is slow, it handpicks a list of articles to narrow its search scope. Due to this, the approach might not find something extraordinary that might have been discovered otherwise. The algorithm uses a user metric scale ranging from 1 to 5 based on coherence, familiarity, and other factors that are very individual-specific and cannot be definitively measured, as they can vary from person to person. It is easy for human minds to find patterns to eastablish pattern between objects even if their is none in the first place, so it may be difficult for metrics to accurately measure the output of two documents that may not be related in reality, but that the human mind associates with a pattern. Informed users may be better able to identify spurious matches and not associate them with new patterns. I do not believe that the idea of a completely closed gap is valid, as stories can link to each other even if they are far apart in time, as seen in an example given by Freakonomics[1], which explores the controversial correlation between legalized abortion and reduced crime rates, with the causal relation happening at least twenty years between the two events. It is also difficult for an algorithm to find such a relation, and any result connecting a set of documents to the crime rate that does not include this relation has no way for the user to verify if the knowledge gap has been closed.
In the second paper, the author attempts to establish a relation between incongruity (this feature contributes the most to humor) based on anomalies and context words. There is no metric upon which the results are based, but conclusions have been made based on the author's observations from a handful of captions in one experiment. The author could have recreated the experiment a few more times with datasets sourced from elsewhere to strengthen their claims, but the nature of the experiment requires a lot of human intervention, which may be a challenge. While this is acceptable, the results cannot be generalized

[1] https://freakonomics.com/2005/05/abortion-and-crime-who-should-you-believe/

because the data source is limited to a few captions (only 16) from one magazine, and it is possible that only a certain section of people read the magazine and submit captions for competitions.

In the third paper, the study uses a clear and standard definition of creativity, including novelty, quality, and feasibility criteria. The study does not account for individual differences in creativity, which may affect the rate at which participants generate good ideas. The evaluation technique focuses primarily on the proportion of good ideas, which may not provide a complete picture of the quality of the generated ideas.

All studies report promising results that could have practical applications in the real world, such as reducing the workload on judges or improving the speed of innovation. Overall, the evaluation criteria in all of the studies were strong, as they modeled the evaluation criteria by breaking it down into multiple performance measures, such as redundancy, relevancy, feasibility, novelty, etc in an attempt to pinpoint the evaluation criteria and reduce the reliance on human perceptions of what these criteria meant. However, the possibility of a dissimilar set of results if evaluated by another set of judges or a different instance of the same type of dataset cannot be ruled out.

## 4. If I were the author

I would have liked to handle a few things from what has been reported above , To better understand humor in jokes, it would be a necessary studt to consider the backgrounds, cultures, and demographics of participants. Select jokes that cater to a diverse audience and gather information about participants then analyze the results more accurately.
I would like to handle punctuation carefully, as it can change the meaning of a joke, so in a way making sure the text processing step takes into account the role of punctuation in the joke.
Instead of only looking at single words, considering n-grams as features for analyzing jokes. This can help capture the context and nuances of humor better.
To address the issue of chronology in the first paper, I would like to develop a method that considers the time of events, even if they were reported later. This can help to capture a more accurate representation of the story.
For the third paper, I would adjust the keyword selection strategy to account for multiple purposes or mechanisms in a text and use the number of keywords (D) proportionally according to the number of purposes and mechanisms present in the text so that a fair distribution can be assumed.

## References


[1] Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 235–243, 2017. 1

[2] Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–632, 2010. 1

[3] Dafna Shahaf, Eric Horvitz, and Robert Mankoff. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1065–1074, 2015. 1