# relevance-paradox

Analyzing trends and phenomena including Simpson's paradox, Berkson's Paradox, and Lord's Paradox in real world Natural Language Data

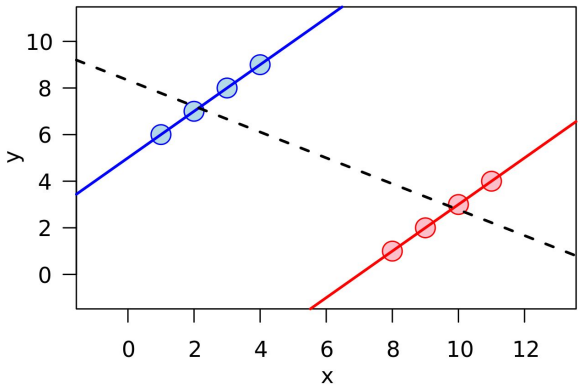
## Introduction

The task of predicting the importance of every sentence in a document on the web can be framed as a regression task where we try to compute a saliency score for every available sentence. Now, we can analyze trends in the sentence-salience data by plotting the normalized relevance score or some other metric against various possible factors such as length of the sentence, the content word count to text length ratio, semantic similarity between the sentence and the document embedding, etc. We can also visualize various other settings and try to specifically find cases of paradox such as Simpson's, Berkson's and Lord's.

## On the Statistical Paradoxes

I'll present the idea behind Simpson's, Berkson's and Lord's paradox here without including extraneous details. Simply observing the mean or correlation between two variables does not necessarily capture the underlying principles within a dataset. For example, if we probabilistically see mortality rate decreasing for smokers in a fairly large population, and we find no correlation between smoking and lung cancer and make the claim that 'smoking saves lives', we would be making a misguided claim. Thus, it would be a case of misuse of statistics where the statistician or the data-scientist does not look into underlying factors, the data environment, unconsidered variables and features and the overall explainability.

When the overall data trend shows a correlation which does not agree with the correlation of it's semantically well defined data clusters, we have a general case of Simpson's paradox. The trend seen in the population is nullified or even reversed within the individual clusters. The most common example here would be that of the treatment process for kidney stones (or smoking and mortality within different age groups). In such cases, the correlation dynamics of the complete data is different from that of its category-based clusters.




Lord's paradox is simply the continuous version of Simpson's paradox where we move from a categorical setting to a continuous one. Lord's paradox examples typically use disjoint clusters to prove a point. Suppose, we consider the same type of value in the x and y axes, i.e the outcome and feature are of the same continuous type. If the same stories of means not conveying the effective measures of central tendency or correlation of individual clusters strongly deviating from that of the complete data are witnesses, we would have a strong case of Lord's paradox. The change in weight based on gender or used dining hall is an excellent illustration in this particular case.[1]

Berkson's paradox is closely related to Simpson's. Two statistical values can appear to have a positive correlation on a large dataset. For example, for all students in a school, we may find a positive correlation between shoe sizes and mathematical ability. This is not explainable or intuitive and the key factor here is the grade or age of the student. A student in a higher grade will be better in mathematics and have a larger shoe size. The same experiment when performed on students of the very same grade shows uncorrelated values.

Thus, the three types of paradoxes are closely tied. In our experiments, we'll study the features and correlations both graphically and statistically.

## Data Collection

The Internet Documents Dataset was prepared in the Information Retrieval & Extraction Lab in collaboration with Microsoft. For news articles on the internet, a single importance score was assigned based on the incoming search queries received by the Microsoft Bing search engine. The semantic overlaps between search queries and sentences in the web-page text was aggregated to a salience score for each sentence. This dataset has been previously used for designing a fact-salience prediction pipeline.

## Data Cleaning

[1] Refer here

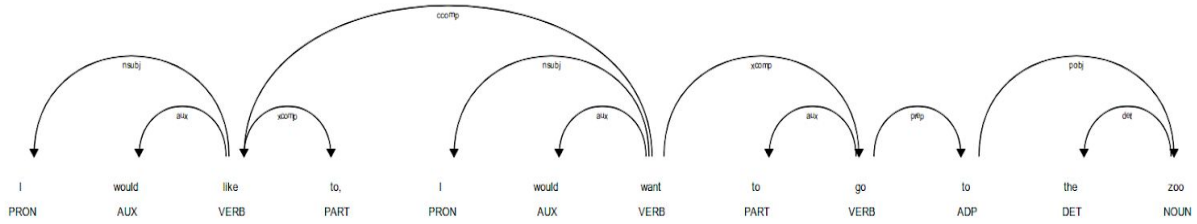
For the purpose of our experiment, extensive data cleaning was required. This is because we only wish to consider proper grammatical sentences with well defined linguistic properties. Also, we cannot have junk text and meta information included in our data as that would upset the overall results. This is essentially the noise attributed to web scraping.

Thus, a lot of further data cleaning had to be done for our existing dataset to be an approximate gold standard for this experiment. Note that there is no well-defined way of achieving this task. We look at the parse structure of the sentences. A dependency parse captures relevant relationships between tokens in a sentence. We use linguistic cues to rule out certain structures.

Let's look at two toy examples:

$S_1$ = "I would like to, I would want to go to the zoo."

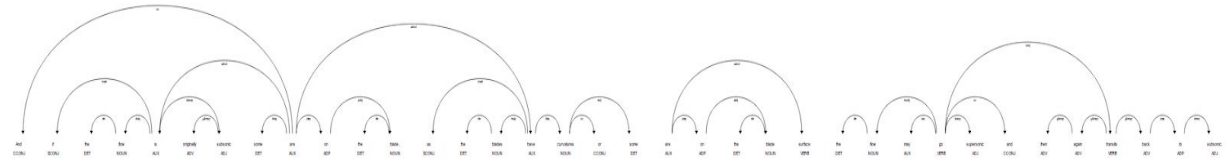
We get the following parse output using Spacy's dependency parser:




The parse structure shows an 'xcomp' branch leading to a single 'to' which is a participle at the base level. This indicates that the phrase is incomplete. An 'xcomp' branch must lead to a complete clause, which is missing in this case. This error originates from the ungrammaticality of the sentence. In "I would like to", the 'to' must lead to a complete clausal complement and the lack of one, results in a violation.

$S_2$ = "And if the flow is originally subsonic some are on the blade, as the blades have curvatures or some are on the blade surface the flow may go supersonic and then again transits back to subsonic."

$S_2$ yields the following parse output:



Clearly, the above dependency tree is not even a connected graph. It is composed of three subtrees. This clearly is an error as a dependency structure is supposed to be a tree. The underlying intuition here is the fact that in order to qualify as a valid sentence, all its individual components must share some sort of connection. Again, the error arises purely due to ungrammaticality.

In our dataset, we have 'sentences' such as:

"*Accounting CS  Current information and alerts Release user bulletins - current and prior WebEx Support Center for CS remote sessions Workflows in Accounting CS Entering data for Form 941 Reporting and Report Designer  Procedures Processing corrected payroll tax forms Troubleshooting an unreconciled amount in bank account reconciliation Address verification tips and tricks Standard reports list  Getting started Finding answers in the Help & How-To Center Accounting CS Ideas Community Transitioning from CSA to Accounting CS Services Getting started with Accounting CS Workflows in Accounting CS Efficiency tips for Accounting CS CS Professional Suite application security overview Contact us Send Support an email message Chat with Support Phone number and queues Support hours Leave feedback Internal Employees Submit H&HTC feedback Submit Video / Service feedback Contact information (optional)  Leave this blank  Please tell us how we can make this information more helpful.*"

This is essentially noise which can be effectively classified and pruned out using the dependency structure violation explained above.

## Experiment

We consider the following features for experimentation purposes:

1. The actual sentence importance score which is part of the cleaned data.

2. Standardized sentence length: A custom built tokenizer was used which considers letters and digits as separate entities, and the number of tokens in each tokenized set was counted. Tokenized sentences of length $> 100$ were ignored and the token-set length was divided by a factor of 100 to bring it to the 0 to 1 range.

3. Content Ratio: The ratio between the number of content tokens in a sentence and its length. The tokens such as nouns, verbs, named entities, etc which carry the weight of the sentence meaning are the considered content words. Deviating slightly from the true linguistic definition, I chose to ignore be-verbs as well.

4. Semantic Similarity: A key idea in text summarization literature is as follows: if the sentence meaning matches the overall meaning of the document such that the sentence can serve as a good representation for - what the document is all about, we can consider the sentence summary worthy. To achieve this, I utilized a Transformer[2] based language model, namely, RoBERTa-large-MNLI[3] which is pretrained using a masked language model objective and fine-tuned on natural language inference data. The embeddings for the sentences and the entire document was computed and the cosine similarity between document and sentence is what was utilized.

Overall, we had over 82,000 web pages and we considered only grammatical sentences. Refer to the Jupyter Notebooks for more experimental details. The flow is fairly easy to follow and self

[2] Attention is all you need
[3] RoBERTa

explanatory. Other utility code has also been provided. The data is not publicly released yet and cannot be shared for that reason.

## Results

We evaluated each of the available possible pairs of features for the (x, y) axes with one of the other features as a control variable. Only one case of paradox was found and for the rest, the overall data correlation matched that of its subparts.

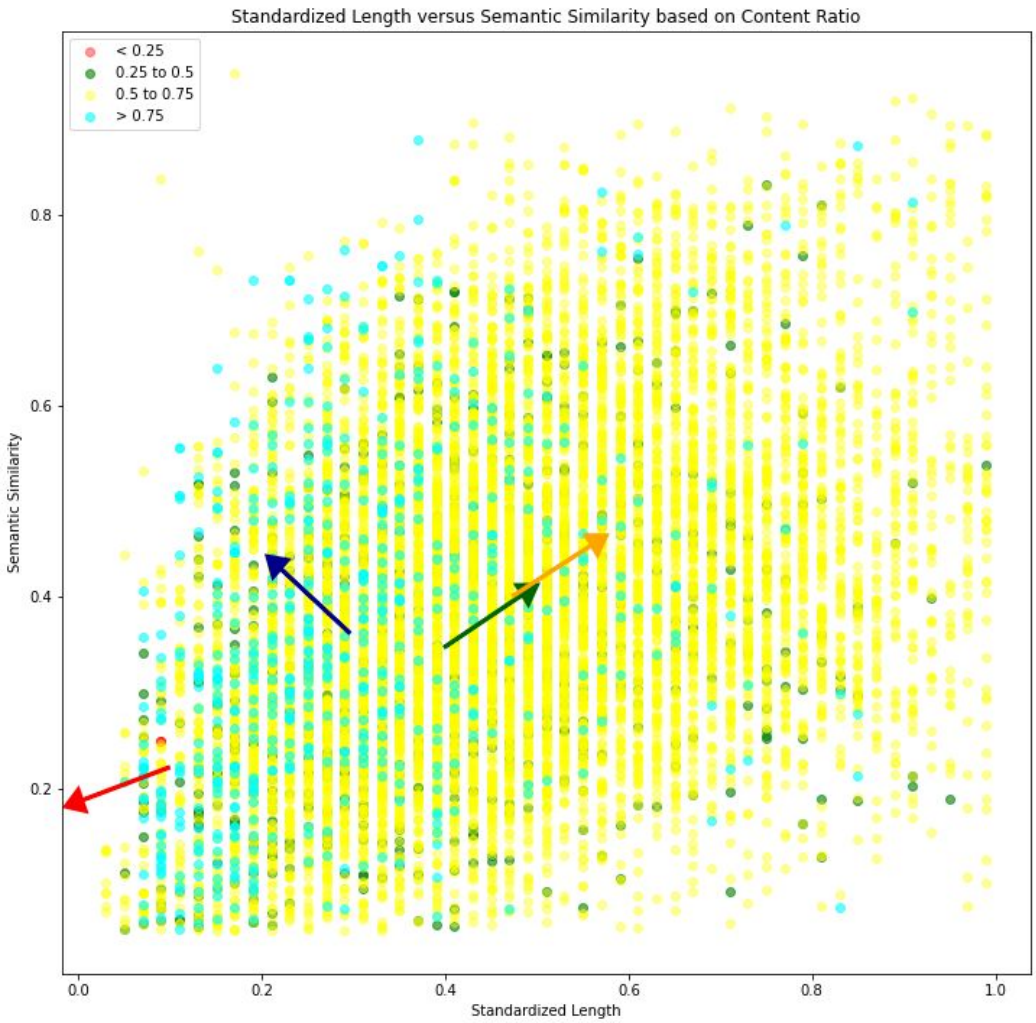
For the control, we divided the range into four disjoint units and viewed the correlation of each part independently. We considered the direction of the principal eigenvector as a measure of correlation.

In all cases except one, the correlation within data clusters agreed with that of the population. Kindly refer to the 'paradox' notebook for more details. However, in the experiment on Semantic Similarity versus Standardized Length, we observe an overall positive correlation within the data indicating that semantic similarity with the document increases with an increase in sentence length, which makes sense. Longer sentences will have more content and be a better representation for the entire document or webpage.

However, we bring in the control factor here which is content ratio. Again, a greater content ratio indicates more information and lesser filler content. We divide the content ratios into 4 categorical ranges:

1. Up To 0.25: Low
2. 0.25 to 0.5: Medium Low
3. 0.5 to 0.75: Medium High
4. Greater than 0.75: High

We observe something interesting here. The overall correlation for the entire dataset resembles that of categories 1, 2, and 3. Note that the direction of the arrow is not important, the general slope of the line is what matters. For category 4 i.e very high values of content ratio, there is a negative correlation between semantic similarity between document and sentence and the overall sentence length.




This is explainable as follows:

Sentences with extreme values of content ratio are mostly informative ones speaking about specific events or activities. It mostly fixates on one very specific activity and focuses on that. For example: In an article that speaks about Daenerys from Game of Thrones, the sentence: “The full title goes like, Daenerys of the House Targaryen, the First of Her Name, The Unburnt, Queen of the Andals, the Rhoynar and the First Men, Queen of Meereen, Khaleesi of the Great Grass Sea, Protector of the Realm, Lady Regent of the Seven Kingdoms, Breaker of Chains and Mother of Dragons.” - will have a very high content ratio value but will not contribute to the overall document meaning to a great extent. These sentences typically fixate on one tributary thought which deviates from the central thought described in the document. Now, if the sentence length increases, we have more content which does not necessarily talk about the central idea of the document and the semantic similarity decreases.

Thus, we have finally found a case for philosophical paradox in natural language data.

The features of semantic similarity and standardized length appear positively correlated in the population. But values within a class (high content ratios) show that the correlation is actually negative for specific segments and we cannot make an overall generalization. This is a fine example of Simpson's and Berkson's paradox. We don't get a proper case of Lord's paradox here because the semantics clusters all overlap and there are no individual disjoint segments.

## References:


1. Simpson's Paradox explained
2. Berkson's Paradox explained
3. Original Description of Lord's Paradox
4. Reversal Paradox
5. Attention is all you need
6. RoBERTa
7. Simpson's Paradox - General Info
8. Berkson's Paradox - General Info
9. Lord's Paradox - General Info