Analysis and Summarization of PubMed Figure-Associated Text and Captions

Tin Oreskovic (to2320), Bindi Patel (bpp2112)

Spring 2018

Abstract

Biomedical literature, including the corpus provided by the US National Library of Medicine National Institutes of Health through PubMed, integrates with the textual information visualizations that provide for a rich source of knowledge themselves. Researchers and individuals seeking to inform themselves with regards to the multifaceted medical and health related issues access these articles and often reference the images or diagrams as well as their captions for a quick synopsis of the larger text as a whole. The captions from a corpus of figures from PubMed articles were analyzed to gain insight on which topics, across PubMed, could be assigned to their contents. The findings are presented in a condensed manner with an interactive visualization tool. The focus was to to see if there are universal methods of describing figure content in their captions.

1 Introduction

Data visualizations such as figures in articles and their respective captions often provide a condensed source of information contained in the larger article as a whole. Often, researchers aim to have these figures as a method of intuitive display of complex analysis; however, this often misses the mark.

Data visualization is thought of as a crucial part of data analysis. Researchers commenting on the dependency of the figures to their corresponding text comment, “By themselves, these figures are nearly always incomprehensible to both humans and machines and their associated texts are therefore essential for full comprehension” [1]. Large amounts of patient data are often summed up in only a few sentences. An analysis of these sentences can allow for insight into the improvement of such figure captions to increase understandability.

Text mining has been conducted by several researchers by applying text classification, terminology extraction, relationship extraction and hypothesis generation on the text of the articles [2]. However, the focus of these various machine learning techniques have not been on the figures included in articles. Similarly, in “Figure-Associated Text Summarization and Evaluation,” [1] researchers comment on how the text associated with figures is not solely in an image’s caption or legend but rather throughout the paper.

In 2006, Yu and Lee [3] hypothesized that “much of image content reported in a full-text article can be summarized by the sentences in the abstract of the article.” They implemented a UI BioEx that allowed biologist to obtain images from abstract sentences and found that over eighty percent of biologists favored this method. Kim utilizes figure text extraction to explore the problem of image captions being incorporated within figures, given the sheer volume of biomedical publications [4] .

In “Are Figure Legends Sufficient? Evaluating the Contribution of Associated Text to Biomedical Figure Comprehension,” [5] researchers had twenty research participants score their level of understanding of figures and associated text. The responses were then evaluated with a ROUGE score.

Existing research qualitatively comments on the understandability of these captions. Research participants provide subjective responses based on their understandability, which may lead to manifestations of underlying biases.

Understanding a universal set of ways figures are described in relative terms and in terms of comparison can serve as the evidence for a new data analysis language. In order to facilitate and promote accurate communication in the biomedical sciences, analysis of figure captions can be used to improve potentially misleading or incomplete captions.

2 Audience and Needs

This project may impact the academic community as well as, potentially, patients who may be researching their ailments through bio-medical journal research articles.

This community is a benefactor because its members often refer to the likes of BioText, which provides a literature search engine that allows biologists to search over the contents of Open Access Journals, and provides figures and respective captions from the articles into the results of a search. By providing an analysis of the topics in a large corpus of the captions, this can potentially lead the way in developing a data analysis language that can eventually improve these captions to provide a more complete summary of the respective image. This document is imagined as first step in this process, since a clear understanding of the present state of the topics that the captions can be assigned to (and from which the captions are generated) is necessary for any improvements.

3 Data Source

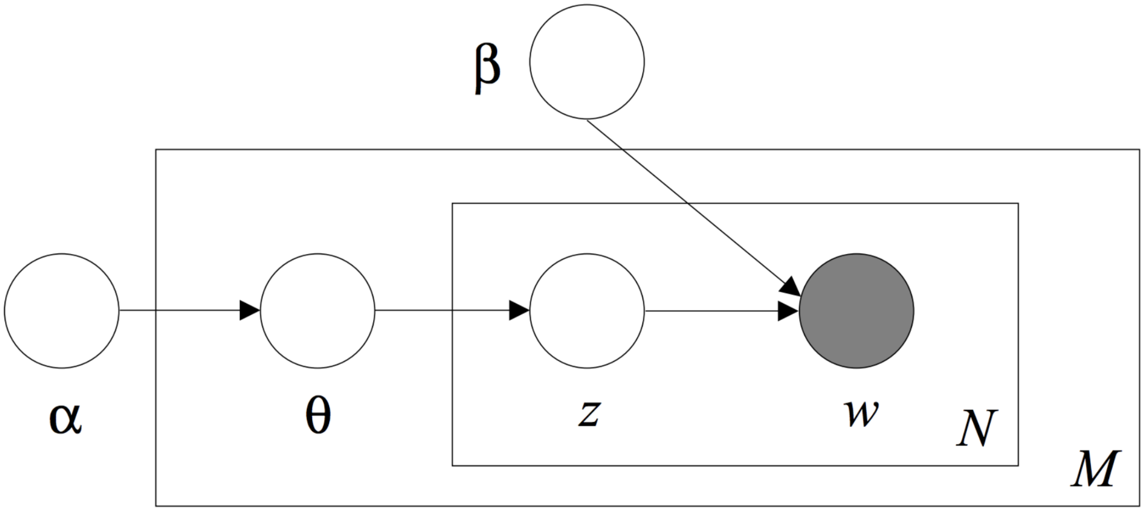
The captions were obtained using the VizioMetrics Open Data API which returns images and their respective captions. [1] A key roadblock to the data collection was that the keywords appearing in captions or paper abstract were required as an input to the API call. To overcome this obstacle, the keywords from the top

[1] link to API documentation

100 articles were aggregated. [2] The corpus of captions used in this analysis is, therefore, limited to the aggregation of the top 1500 articles that are produced from the API call respective to each of the keywords.

4 Approach

To analyze the corpus of figure captions and extract the topics that the text is about we implemented the Latent Dirichlet Allocation (LDA) - a generative statistical topic model that aims to explain the linguistic content of documents by estimating which are the topics that each document consists of. Each document (here, each caption in the corpus) is assumed to be composed of the extracted topics, where each topic is characterized by a multinomial probability distribution over all the phrases. LDA operates by assuming a model that can be represented graphically (exactly as in the original paper [6]):



Here, α is a parameter of a "Dirichlet prior" specifying the prior beliefs about the topic distributions for each caption, which we leave unspecified in our implementation due to a lack of a priori information, except for the below-explained tacit assumption about the number of topics, which is required; β is the parameter of the corresponding Dirichlet prior capturing any prior beliefs about the multinomial distribution of two-word-phrases within a given topic; θ_m is the topic distribution specifying which topics determine the phrases it consists of, for document m; φ_k is the phrase distribution for topic k, from which, according to the model's assumptions, the words are drawn randomly to compose the documents along with words from the other topics assigned to the document; z_{mn} is the topic that the n-th phrase belongs to in document m; finally, w_{mn} is the particular phrase. The code used for our implementation of LDA as well as for the below-described visualization tool is available here.

It should be noted that, thus, the topics extracted by an algorithm implementing an LDA model do not precisely correspond to the popular intuitions about "topics": the phrase-content (i.e. text) of each caption is assumed to be determined by the topics it is assigned and the phrases that each of these topics consists of, according to the corresponding probability distributions. We set the number of topics to 4, as this is the largest number such that the Jensen-Shannon divergence between more than two topics isn't very small (indicating that they are very similar, which may suggest that a higher interpretability is achievable with fewer topics). Furthermore, we choose to focus on two-word-phrases instead of the "unigrams" (i.e. single words) discussed in the original paper, because, after testing both variants, we concluded that the interpretability is higher when considering phrases, which convey richer information than single words in this

[2] link to PubMed Trending Articles

highly specialized context: the technical meaning of “confidence interval,” for instance, will be separated from that of “high confidence.”

Then, with the output we obtained, we visualized the findings by with the below interactive tool, originally developed and studied by Shirley and Sievert, to aid answering three questions:

- What is the meaning of each topic?
- How prevalent is each topic?
- How do the topics relate to each other? Different elements of the visualization address each of these questions.

The left and right panels of our visualization are connected so that clicking on a topic (circle) on the left renders the most useful (explained below) two-word phrases for interpreting the selected topic. To answer the second and third questions, one uses the panel on the left: the area of a circle is proportional to the overall prevalence of the corresponding topic in the town hall transcript, while the distance between the centers of two circles is here shown as proportional to the (symmetric) Jensen-Shannon divergence between any two topic probability distributions - in other words, the further away two circles are, the less similar the topics they stand for are.

Hovering over one of the phrases in the horizontal bar-chart on the right panel once a topic is already selected yields the conditional distribution over topics (on the left) for the selected phrase, with the changing circle areas emphasizing the conditional probabilities of that phrase appearing within other topics where the phrase is present (and the remaining topic circles disappearing). This is meant to allow exploring the relation between topics and phrases, but it also adds to the intuitive experience of the distance between the topics.

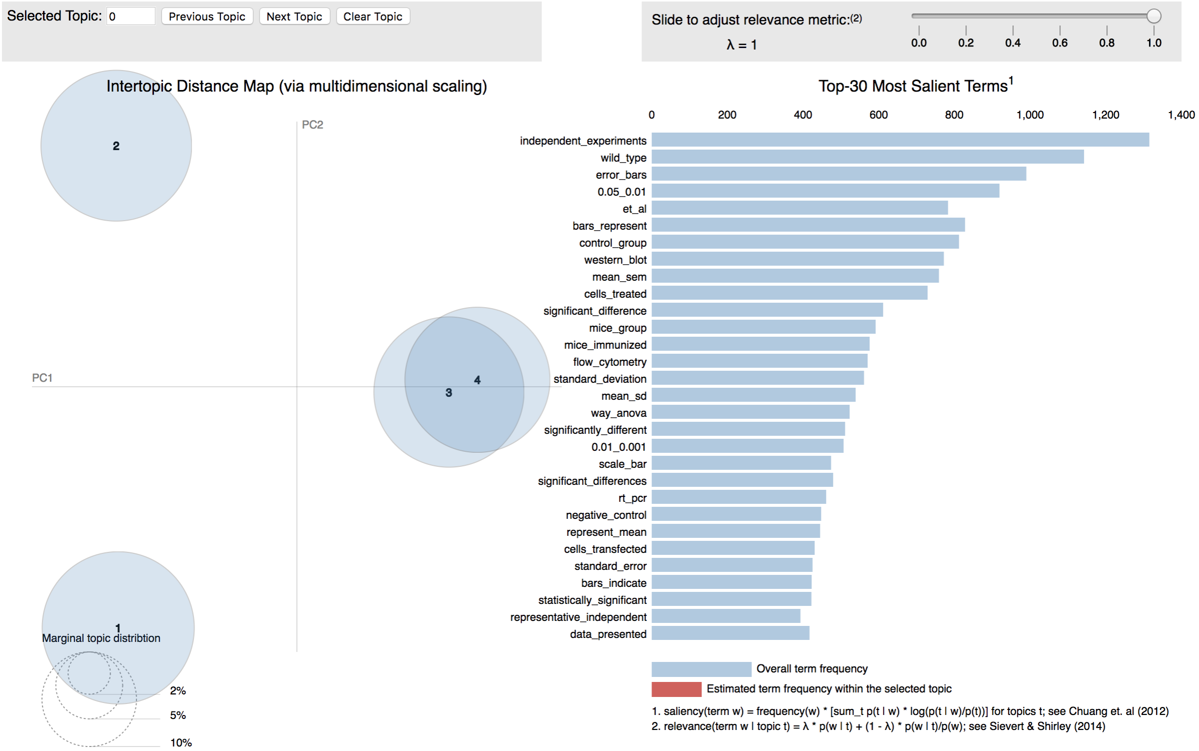
Finally, “to interpret a topic, one typically examines a ranked list of the most probable terms in that topic, using anywhere from three to thirty terms in the list. The problem with interpreting topics this way is that common terms in the [whole] corpus often appear near the top of such lists for multiple topics, making it hard to differentiate the meanings of these topics.” [7] For this purpose, included here is the interactive λ parameter as a part of a relevance metric, where

$$relevance(ph_i, top_k \mid \lambda) = \lambda \log(\phi_{phr_i, top_k} + (1-\lambda) \log\left(\frac{\phi_{phr_i, top_k}}{p_{phr_i}}\right).$$

In accordance with the paper by Sievert and Shirley, ϕ_{phr_i, top_k} is here the probability of the phrase i in topic k, while p_{phr_i} is the marginal probability of the phrase i in the entire corpus of captions. In short, “setting $\lambda = 1$ results in the familiar ranking of [phrases] in decreasing order of their topic-specific probability, and setting $\lambda = 0$ [phrases] solely by their” topic specific probability divided by the overall marginal probability of the phrase in the whole transcript (for details, see the original paper). In other words, by setting the λ one decides how much to weigh more uniquely characteristic phrases in the ranking for a given topic rather than just by ranking the straightforwardly most frequent phrases within the topic, which may be very frequent across the town hall transcript. The optimal value is subjective and depends solely on what is most suitable for human interpretation of the meaning of a topic. By examining the resulting bar charts in the right panel, one can hopefully infer what the topic is about.

5 Results

Below is a static image of the interactive visualization, since the latter cannot be included in pdf files. The interactive visualization, which is in fact referenced to above and used for the inferences about the topics below is available here.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 1: Please access the actual interactive visualization of the output of the implemented LDA here.

We address the first question the interactive tool was developed for by examining the bar charts while setting somewhat arbitrarily around 0.5 for a ranking balance between more unique yet still very frequent phrases within each topic. In this particular case, the exact value of λ does not greatly impact the ranking of phrases rendered in the horizontal bar charts, which indicates that the output of the LDA, when considered even by the straightforward ranking of the most frequently appearing phrases within each topic, appears to produce highly distinctive phrases across topics. The meanings of the four extracted topics are (subjectively) described below:

1. The focus of this topic is on statistical methods and analysis. The images relevant to captions for which this topic was an important component are likely visualizations of comparisons to a known baseline or a control group. The captions, then, are probably related to analyses of variations and differences between these groups, such as with Analysis of variance (ANOVA).

2. The captions in this topic are domain specific to molecular biology. Moreover, the focus appears to be surrounding research methods to analyze and improve treatment. Techniques such as blot analysis are likely visualized in the corresponding images and diagrams. Salient terms include “cells infected” and “cells treated.”

3. This topic appears to be focused on clinical trials. Terms include “treatment group,” “time point,” “positive control,” and “significantly higher,” which are comparative phrases. Kaplan-Meier curves are likely visualized with certain captions in for which this topic is an important contributor. Confidence interval(s) is another salient term. The captions may include population estimates from the trials’ findings.

4. This topic contains descriptive captions. Comparative phrases such as “independent experiments,” “0.05 versus,” and “control groups” appear to be among the most relevant phrases. Other salient

phrases highlight various aspects of the corresponding image; for example, "original magnification" and "figure shows" are explanatory phrases clarifying the accompanying diagram or image.

All the topics are represented as circles of similar areas, indicating that their prevalence was similar in the corpus, i.e. the set of considered figure captions. The distance between the distributions of the topics (from one circle's center to that of another) is smallest between (3) and (4), which corresponds to the prominence of comparative phrases in both. Hovering over those phrases will reveal their conditional distribution across topics, with the other of the two topics featuring prominently. The distributions of topics (1) and (2) are approximately equidistant from (3) and (4), which again supports (and drives) our interpretations of their content as more distinctive relative to the remaining topics than that of (3) or (4).

6 Conclusion

Across topics which may be more domain specific or more grounded in analysis, there is an underlying comparative nature to the captions. The salient phrases highlight this functional feature of the captions across the corpus of PubMed articles analyzed. Captions are used to provide context by comparing to some known theoretical, conjectural, or true baseline. These comparisons are expressed in terms of both qualitative terms describing the associated figure and quantitative, statistical summaries of analysis conducted.

7 Future Work

The ideal impact resulting from this project would be a corpus of common text or a universal set of phrases that embodies the current state of PubMed captions. A wider net may be cast to incorporate less common keywords to provide a more complete analysis. The captions corresponding to different domains can be analyzed independently to see if there is domain specific variation. These further steps may yield a more nuanced and specific analysis.

References

[1] Ricky J. Sethi Polepalli Ramesh, Balaji and Hong Yu. Figure-associated text summarization and evaluation. *PLoS ONE* , 10.2, 2015.

[2] WR Hersh Aaron Cohen. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6, 2005.

[3] Minsuk Lee Hong Yu. Accessing bioscience images from abstract sentences. *Bioinformatics*, 22, 2006.

[4] Hong Yu Kim D. Figure text extraction in biomedical literature. *PLoS ONE* , 2011.

[5] Hong Yu. Are figure legends sufficient? evaluating the contribution of associated text to biomedical figure comprehension. *Journal of Biomedical Discovery and Collaboration*, 4, 2009.

[6] Andrew Y. Ng David M. Blei and Michael Jordan. Latent dirchlet allocation. *Journal of Machine Learning* , 3, 2003.

[7] Kenneth E. Shirley Carson Sievert. Ldavis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* , pages 63–70, 2014.