Explanation Mining

Bhavya
University of Illinois at Urbana-Champaign
bhavya2@illinois.edu

ChengXiang Zhai
University of Illinois at Urbana-Champaign
czhai@illinois.edu

ABSTRACT

Explanations are used to provide an understanding of a concept, procedure, or reasoning to others. Although explanations are present online ubiquitously within textbooks, discussion forums, and many more, there is no way to mine them automatically to assist learners in seeking an explanation. To address this problem, we propose the task of Explanation Mining. To mine explanations of educational concepts, we propose a baseline approach based on the Language Modeling approach of information retrieval. Preliminary results suggest that incorporating knowledge from a model trained on the ELI5 (Explain Like I'm Five) dataset in the form of a document prior helps increase the performance of a standard retrieval model. This is encouraging because our method requires minimal in-domain supervision, as a result, it can be deployed for multiple online courses. We also suggest some interesting future work in the computational analysis of explanations.

Author Keywords

Explanations; Language Modeling for Information Retrieval; Prior Probability

CCS Concepts

•**Applied computing** → **Computer-assisted instruction;**

INTRODUCTION

The quote "Good teaching is good explanation" (Calfee 1986: 1-2) captures the indispensable role of explanations in teaching. Leinhart [5] define *instructional explanations* as explanations that aim to explain concepts, procedures, etc. in order to help students understand and use information in a flexible way. In the present age, one of the main ways of learning is by reading text materials available online in various forms such as textbooks, research papers, lecture slides, and many more. During this process, learners may encounter certain text segments that they do not fully comprehend and might need to find their explanations. These text segments could range from small units such as phrases (typically corresponding to concepts) to larger units such as a paragraph or a slide. Currently, the only way to find explanations is by using search engines that may not work well as they are not specialized for this purpose. Moreover, search engines typically retrieve documents and not excerpts. To overcome these challenges, we propose the task of Explanation Mining. The research question we address is how to extract suitable explanations from rich online educational sources, such as textbooks.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
L@S '20, August 12–14, 2020, Virtual Event, USA.
© 2020 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-7951-9/20/08 ...$15.00.
http://dx.doi.org/10.1145/3386527.3406738

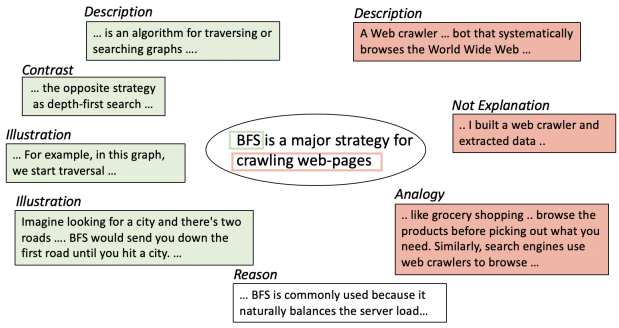

Figure 1. Illustration of Explanation Mining subtasks. Sections in green are candidate explanations for "BFS" and those in orange are for "crawling web pages". The "Reason" at the bottom is the explanation for the implied question "Why is BFS a major crawling strategy?"

Our task is related to early work on Explanation Systems [7, 8] and Adaptive Educational Hypermedia [1] to assist users with learning and understanding. However, these systems require extensive manual knowledge engineering efforts and thus are not scalable or generalizable. There is also some limited work on generating template-based explanations for specific domains [3]. We aim to develop general techniques.

PROBLEM DEFINITION

Given an inquiry q, the goal is to find explanation(s) $(x_1, .., x_n)$ that will help increase a user's understanding of concepts in q. Explanation consists of a list of sentences $\{s_0, s_1, .., s_n\}$ that can be extracted from some source text t_e. To allow personalization and contextual explanation mining, there might also be additional context variables indicating the user's knowledge level or subject/domain.

SUBTASKS

We identify the following subtasks of Explanation Mining:

Identifying units requiring an explanation

Text segments or units that need to be explained could range from small units such as phrases (typically corresponding to concepts) to larger units such as a paragraph or a lecture slide. If the inquiry has several concepts, it becomes challenging to identify which concepts and/or the interactions between them require an explanation. For example, if the user seeks to understand the statement "Breadth First Search is a major strategy used for crawling webpages", there can be 3 underlying questions or gaps in user's understanding: (a) "What is Breadth First Search?", (b) "What is webpage crawling", and

(c) "Why is Breadth First Search a major crawling strategy" (See Figure 1). Further, the system may order these questions based on their dependencies, e.g. a basic understanding of (a) and (b) is pre-requisite to the understanding of (c). In this way, multiple hierarchical explanations may be mined and presented to users.

Explanation Classification/Ranking

Given any piece of text, the goal of Explanation Classification is to identify whether it is an explanation or not. Here, we may assume that we have pre-extracted segments of text from t_e. We may additionally detect the boundaries of explanatory segments within t_e.

Alternatively, we may also pose a ranking task where *better* explanations are ranked higher. Given two *correct* explanations, one may not be universally considered to be better than another. However, we can identify good explanations for a fixed context. For example, if the goal is to explain "BFS" to a layperson, a simple explanation in plain language is more preferable (e.g. the two Illustrations for "BFS" in Figure 1).

Identfying components and types of explanations

There are two major constituents of an explanation: the explanandum (the text describing the phenomenon to be explained) and the explanans (the statements constituting the explanation) [4]. Detecting these constituents is useful to identify whether a given explanation explains the inquiry, i.e. whether explanandum is a part of the inquiry. To select the most suitable explanations for a given context, it is also useful to categorize explanations based on their explanation techniques, i.e. Illustrations, Description, etc. (see Figure 1).

MINING EXPLANATIONS OF EDUCATIONAL CONCEPTS

We will now discuss a specific instance of explanation mining for finding explanations of educational concepts. In particular, given a concept as an inquiry q , the goal is to retrieve a ranked list of explanation units or text segments $x = (x_1, .., x_n)$ from a collection C. Here, we assume a fixed context so that the goodness of an explanation is fixed. Further, we assume that the candidate explanation units are pre-extracted.

Method

We use the Language Modeling approach [10] for retrieval as it is one of the most popular unsupervised approaches. This is based on the Bayes' formula for estimating $p(x|q)$, i.e. probability of generating a explanation x given a query q. Concretely, $p(x|q) \propto p(q|x)p(x)$. Here, $p(x)$ is the prior probability of a candidate unit being a good explanation to *any* query and $p(q|x)$ is the query likelihood given the explanation.

For estimating $p(x)$, we use the ELI5 (Explain Like I'm Five) subreddit [1] where users ask questions on multiple topics (e.g. Biology) and request for layperson-friendly explanations. Each question in the ELI5 forum can have multiple explanations which are scored by users. We can consider the score of an explanation as a proxy to its goodness and thus, automatically obtain a training set for training a supervised Learning-to-Rank (LTR) model [2]. Since our goal is to identify good explanations independent of the query, we use query-independent features such as ngrams, length, readability score of explanations. This pre-trained model is used to estimate $p(x)$. Let $p_{ltr}(x)$ be the probability of x being a good explanation obtained from the LTR model. Since $\sum_{x \in C} p(x) = 1$, we have:

$$p(x) = \frac{p_{ltr}(x)}{\sum_{x \in C} p_{ltr}(x)} \quad (1)$$

We also want to study and fine-tune the contribution of the prior towards the final ranking. Therefore, we perform a linear interpolation of the query log likelihood and the log of the prior as follows:

$$\log(p(x|q)) = (1-\alpha)\ \log(p(q|x)) + \alpha\ \log(p(x)), \quad (2)$$

where α is a parameter that can be tuned.

For estimating $p(q|x)$, we use Language Modeling with Dirichlet Prior smoothing as it is known to perform well for short queries [9] like concepts which are generally noun-phrases.

EXPERIMENTS

Dataset

Textbooks are excellent sources of explanations of concepts. Therefore, we use a textbook on Text Mining and Information Retrieval [10] to create our dataset. This textbook is intended for upper-level undergraduate students. As a basic approach to obtain explanation units, we parsed the book using GROBID [2] and treated each extracted section as a unit.

To obtain queries automatically, we leveraged the index section of the textbook. The index contains a list of concepts and the page numbers where the explanations of those concepts can be found. Thus, concepts are used as queries and the sections on the corresponding page numbers as the relevant explanation units. We obtained a total of 248 queries and 323 relevance judgments. Generally, there are 1-2 relevant sections per query.

Setup

We are interested in investigating two questions: Does incorporating the prior using our interpolation method help increase the performance compared to Dirichlet Smoothing alone? How does the performance vary with the parameters α and the smoothing parameter in Dirichlet Smoothing (μ)?

For the first question, we performed 5-fold cross-validation using an 80-20 data split. We tuned μ in the range [0,5000] in increments of 500. For each μ, α is tuned in the range [0,1] with increments of 0.01. We will refer to this set of experiments as **Exp1**. The performance is optimized for NDCG@3 since generally there are only 1-2 relevant sections per query, so we want good performance within the top 3 ranks. We also report MAP@3 and Recall@3 for the test sets. Note that in our setting, recall measures the completeness of an explanation. This is because there is generally only one explanation (continuous span of text) of a concept within the textbook but due to our current explanation unit extraction method, it may be divided into multiple consecutive sections. We mark all such sections as equally relevant. Thus, the ranker should retrieve as many sections as possible for the explanation to be complete and useful. It may be possible to pre-extract complete explanation units but we leave this task for future work.

[1] https://www.reddit.com/r/explainlikeimfive/

[2] https://github.com/kermitt2/grobid

Method	NDCG@3	MAP@3	Recall@3
Dir. Prior	0.6787	0.6460	0.7069
Ours	**0.6840**	**0.6491**	**0.7211**

Table 1. Results of 5 fold cross-validation

For the second question, we vary α at different values of μ on the entire dataset and measure the performance (NDCG@3, Recall@3, and MAP@3). The range of μ and α is the same as in the first set of experiments. We will refer to this set of experiments as **Exp2**.

Implementation details

By performing multiple cross-topic experiments on the ELI5 dataset, we found that a model trained on reasonably large datasets using only ngrams as features has the best performance on ranking ELI5 explanations. Thus, we use a model trained using ngrams on 23k questions collected from July 2011 to 2019. We used LambdaMart [2] for training since it is one of the best Learning to Rank methods.

We use the Meta toolkit [6] for indexing and retrieval as it allows us to easily extend its ranking methods and implement our interpolation-based ranker.

Results

Exp1: Table 1 shows the average NDCG@3, Recall@3, and MAP@3 on the test sets of 5-fold cross-validation. We can see that our interpolation method performs better than Dirichlet Prior alone. This is encouraging as it indicates that even a model trained on a general-purpose dataset of explanations for laypersons helps achieve better performance on retrieving explanations meant for Computer Science/Information science undergraduate students. This also suggests that perhaps there are some markers of explanations that are quite robust and generally shared across different knowledge levels and subjects. It would be interesting to train models on similar domains and grade levels in the future and compare the performances.

The maximum improvement is obtained on Recall@3 (+2% relative improvement) and improvements on NDCG@3 and MAP@3 are marginal (+0.78% and +0.48% relative improvements respectively). The increase in Recall might be because the prior estimate helps in retrieving those sections where the concepts are not mentioned too often, but the sections are good explanations. In such cases, a high prior estimate boosts the overall score compared to query likelihood alone. Precision is improved in a few cases which could be those where the sections contain the words in the query, but they are in fact not explanations (e.g. could be exercise questions). In such cases, the low prior estimate results in a low overall score compared to using query likelihood alone.

Exp2: Figures 2 and 3 show the effect of varying α at different levels of μ on NDCG@3 and Recall@3 respectively. We don't show the performance for $\mu = 0$, 500 due to space limitations, and the overall trends being similar to the performance at $\mu = 1000$. Also, the MAP@3 curves are quite similar to those of NDCG@3 and thus are not shown here again. We make the following observations:

First, we can see that the performance of our interpolation method is generally higher than that of the Dirichlet prior

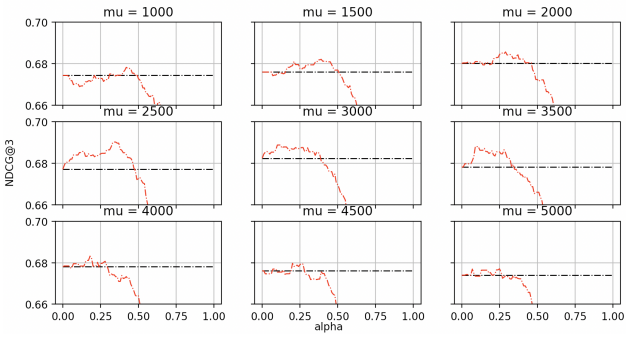

Figure 2. Effect of varying α on NDCG@3 at different values of μ. Black dotted line is the Dirichlet Prior method alone and red dotted line is our interpolation method.

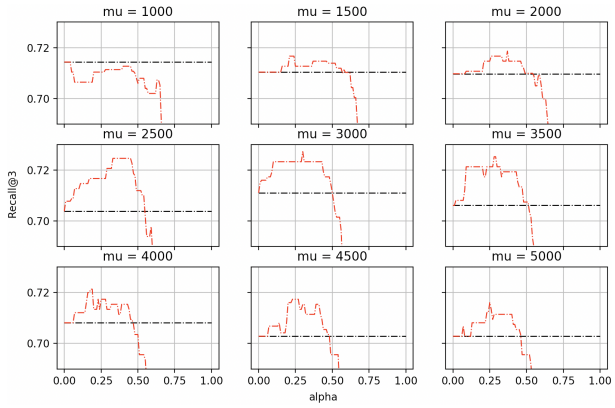

Figure 3. Effect of varying α on Recall@3 at different values of μ. Black dotted line is the Dirichlet Prior method alone and red dotted line is our interpolation method.

method alone for $\alpha <= 0.5$ and then drops, meaning that the query likelihood should be weighed higher than the explanation prior probability. This is expected because the retrieved explanation unit should at least contain the words in the query and not simply be *any* good explanation.

Secondly, the improvement in the recall is more substantial compared to improvements at NDCG and MAP for almost every μ, which is similar to the overall behavior we noticed in **Exp1**. The recall is also less sensitive to the choice of α (has broader curves) than MAP or NDCG.

Table 2 shows the performance on the full dataset using the best run values of the two methods optimized over NDCG@3. Again, we can see that all values are higher using our interpolation method. Improvement in Recall@3 is statistically significant ($p<0.05$) based on Wilcoxon signed-rank test.

Method	NDCG	MAP	Recall
Dir. Prior (μ=3000)	0.6824	0.6491	0.7111
Ours (μ=2500, α=0.34)	**0.6902**	**0.6555**	**0.7246***

Table 2. Results (at 3) of best-run of each method on full data. Best parameters are shown in paranthesis. * indicates statistical significance ($p<0.05$) based on Wilcoxon signed-rank test

FUTURE WORK

From a pedagogical standpoint, there are two kinds of users involved in the process of explaining: explainers (providers of explanation) and explainees (recipients of explanation). In this paper, we focused on Explanation Mining techniques to assist explainees. However, we believe that computational analysis of explanations can assist both these users in the future as discussed below.

Explainee Assistant

The goal of an explainee assistant is to help users find a suitable explanation. Explanation Mining techniques discussed in this paper are fundamental to the development of such systems. Generally, we need to consider the following two dimensions while developing such systems.

Mode of Inquiry: Users may seek an explanation explicitly or implicitly. By explicit inquiry, we mean that users pose a direct question seeking an explanation. An implicit inquiry is when users reading a piece of text (or even watching a video/graph/image) need assistance with understanding it. In other words, users pose an indirect, broad question i.e. 'What does this paragraph mean?'. In this case, the system might need to additionally identify units within the selected text that require an explanation.

Mode of Explanation: Several aspects need to be considered while presenting an explanation to a user. The system can provide explanations in the form of text, videos, images, or a combination of them. The explanation might be directly extracted directly or synthesized from one or more sources. Further, an ideal explanation should be suitable for a learner, i.e. suitable for their knowledge level (e.g. laymen vs undergraduate) and be presented in a way that is interesting to them (e.g. by illustrating in terms of familiar ideas). Finally, the process of explaining should be interactive, allowing users to seek more information until satisfied.

Explainer Assistant

From the perspective of an explainer, it is useful to have a writing assistant that helps them perfect their explanation. The assistant can analyze a given explanation and provide feedback to the explainer through scores or provide suggestions for improvement. As explanations are meant for explainees, the system should consider variables such as their knowledge level and interest while providing the feedback.

Finally, both these application systems can mutually benefit each other. The explainer assistant can help in customizing an existing explanation for a particular explainee. On the other hand, the data collected from the explainee assistant can help in identifying the types and features of explanations preferred by different explainees which can, in turn, be used to train the explainer assistant.

Further research is required on designing these systems and developing various NLP and data mining techniques for mining/synthesizing suitable explanations.

CONCLUSION

In this paper, we present our vision of analyzing explanations computationally for pedagogy. We are motivated by its practical applications to assist both explainees and explainers. Our current focus is on the task of Explanation Mining to automatically find suitable text-based explanations of inquiries posed by explainees. We proposed a basic approach to mine explanations for educational concepts from textbooks based on the popular Language Modeling approach and using an explanation prior. Preliminary results suggest that our proposed method has a better performance compared to using query likelihood alone. This suggests that the same explanation markers that are useful in ranking layperson explanations can help boost performance in specific educational domains as well. Thus, our method can potentially be deployed for many online courses as it does not require much in-domain training. In the future, it will be useful to develop both explainee and explainer assistants and test the utility of our proposed techniques. Since both these application systems are mutually beneficial, a unified framework of explanations that assists both explainers and explainees should be developed.

REFERENCES

[1] Paul De Bra and Licia Calvi. 1998. AHA! An open adaptive hypermedia architecture. *New Review of Hypermedia and Multimedia* 4, 1 (1998), 115–139.

[2] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23-581 (2010), 81.

[3] Andrew Head, Codanda Appachu, Marti A Hearst, and Björn Hartmann. 2015. Tutorons: Generating context-relevant, on-demand explanations and demonstrations of online code. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 3–12.

[4] Carl G Hempel and Paul Oppenheim. 1948. Studies in the Logic of Explanation. *Philosophy of science* 15, 2 (1948), 135–175.

[5] Gaea Leinhardt. 2001. Instructional explanations: A commonplace for teaching and location for contrast. *Handbook of research on teaching* 4 (2001), 333–357.

[6] Sean Massung, Chase Geigle, and ChengXiang Zhai. 2016. MeTA: A Unified Toolkit for Text Retrieval and Analysis. In *Proceedings of ACL-2016 System Demonstrations*. Association for Computational Linguistics, Berlin, Germany, 91–96.

[7] Edward Shortliffe. 2012. *Computer-based medical consultations: MYCIN*. Vol. 2. Elsevier.

[8] William R Swartout. 1983. XPLAIN: A system for creating and explaining expert consulting programs. *Artificial intelligence* 21, 3 (1983), 285–325.

[9] Chengxiang Zhai and John Lafferty. 2017. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 268–276.

[10] ChengXiang Zhai and Sean Massung. 2016. *Text data management and analysis: a practical introduction to information retrieval and text mining*. Association for Computing Machinery and Morgan & Claypool.