An Investigation into the Pedagogical Features of Documents

Emily Sheng, **Prem Natarajan**, **Jonathan Gordon**, and **Gully Burns**
USC Information Sciences Institute
Marina del Rey, CA, USA
{ewsheng, pnataraj, jgordon, burns}@isi.edu

Abstract

Characterizing the content of a technical document in terms of its learning utility can be useful for applications related to education, such as generating reading lists from large collections of documents. We refer to this learning utility as the "pedagogical value" of the document to the learner. While pedagogical value is an important concept that has been studied extensively within the education domain, there has been little work exploring it from a computational, i.e., natural language processing (NLP), perspective. To allow a computational exploration of this concept, we introduce the notion of "pedagogical roles" of documents (e.g., *Tutorial* and *Survey*) as an intermediary component for the study of pedagogical value. Given the lack of available corpora for our exploration, we create the first annotated corpus of pedagogical roles and use it to test baseline techniques for automatic prediction of such roles.

1 Introduction

We define "pedagogical value" as the estimate of how useful a document is to an individual who seeks to learn about specific concepts described in the document. A computational task that operationalizes the concept of pedagogical value is generating an ordered reading list of documents that a learner can traverse in order to maximize understanding of a subject. When a professor manually constructs a reading list about a specific subject for a student, the professor incorporates substantial knowledge of the subject history and interdependencies with other related subjects. The student's background and the relative qualities of documents on similar subjects are also considered. Techniques for automatically generating reading lists should also consider the extent to which a learner will be able to learn from a particular document.

Previously, Tang and McCalla (2009) have studied the "pedagogical value of papers" in the context of paper recommendation. In their work, they define the multiple "pedagogical values" of a paper as the paper's overall ratings, popularity, degree of peer recommendation, learner gain in new knowledge, learner interest, and learner background knowledge. Other efforts on generating reading lists and document recommendation have focused on modeling concepts represented in documents (Jardine, 2014), modeling concept dependencies (Gordon et al., 2016), and user modeling (Bollacker et al., 1999), but there appears to be very limited work on characterizing the learning utility between a learner and a document. The abstract nature of pedagogical value motivates us to identify explicit document features that are salient to pedagogical value. With graduate students as our target learners, we start with a simplified model of novice, intermediate, and advanced learners, and we focus on identifying pedagogical features of documents that could benefit different learners.

In our document annotation process, we collected annotations for the qualitative and largely objective judgments of categories that documents belong to: *Tutorial*, *Survey*, *Software Manual*, *Resource*, *Reference Work*, *Empirical Results*, and *Other*. We identify the seven categories based on document objectives in presenting content, e.g., *Tutorials* teach the reader step by step how to do something, *Resource* papers point the reader to datasets and implementations. Motivated by the need to conceptually organize information to be pedagogically useful, we refer to documents with different objectives as fulfilling different "pedagogical roles." In the rest of this paper, we will use the document category names to refer to the pedagogical roles.

Identifying important qualitative features of pedagogical value, such as the pedagogical role, gives a greater degree of interoperability and insight into how we can help students learn more effectively. Education research explains the distinction between declarative and functioning knowledge: the former is knowledge of content and the latter is knowledge of how to interpret and put the content to work (Biggs, 2011). To apply content, learners must first understand the content; this explains why a novice and an advanced learner trying to learn the same subject would seek out documents with different pedagogical roles. *Tutorials*, *Reference Works*, and *Survey* papers are better introductions for a novice with no knowledge of a subject. In contrast, an expert would have enough background knowledge to dive right into advanced papers presenting state-of-the-art empirical results. Although pedagogical roles are not the same as pedagogical value, these pedagogical features offer some insight as a starting point for estimating learning utility. For our study, we collected annotations for over 1000 documents, which we document and make available for others to use.[1]

We also collected annotations for three ordinal-scale questions of document complexity and quality as an exercise to gauge the feasibility of the task despite its subjective nature. However, the resulting inter-annotator agreement results were too low to be meaningful. These results stress the importance of identifying more objective user and document features relevant to pedagogical value; in this initial work, we focus on document features.

Our contribution is twofold: We provide the first annotated corpus of pedagogical roles for the study of pedagogical value, and we present baseline classification results using state-of-the-art techniques for others to work with. Our goal is to establish a framework that can be extended to other domains, provide empirical results to validate our dataset and algorithms, and demonstrate the feasibility of the proposed role classification task. In the rest of this paper, we will describe our methods for collecting, evaluating, and automatically generating annotations in Section 2, the results of our evaluations in Section 3, related work in Section 4, and concluding remarks in Section 5.

[1] https://doi.org/10.6084/m9.figshare.5202424

2 Methods

2.1 Creating guidelines for annotation

We performed a few rounds of annotation to develop a set of roles that would be adequate and insightful for an initial investigation. We identified the following pedagogical roles:

- ***Survey*:** Is this document a broad survey? A broad survey examines or compares across a broad concept.
- ***Tutorial*:** Is this document a tutorial? *Tutorials* describe a coherent process about how to use tools or understand a concept, and teach by example.
- ***Resource*:** Does this document describe the authors' implementation of a system, corpus, or other resource that has been distributed (e.g., public data sets or tools that have been released under an open-source license or are commercially available)?
- ***Reference Work*:** Is this document a collection of authoritative facts intended for others to refer to? Reports of novel, experimental results are not authoritative facts; the statement "grass is green" is. *Reference Works* describe different subtopics within a concept.
- ***Empirical Results*:** Does this document describe results of the authors' experiments?
- ***Software Manual*:** Is this document a manual describing how to use different components of a software?
- ***Other*:** Other role. This includes theoretical papers, papers that present a rebuttal for a claim, thought experiments, etc.

Additionally, we developed annotation guidelines instructing annotators to select all applicable pedagogical roles for each document. A document could present results of a novel method and also direct readers to an implementation of the method, thus making the paper both an *Empirical Results* paper and a *Resource* paper. Another document could simultaneously give a step-by-step tutorial about how to use a system, present specific commands on how to use components of the system, and provide a link to where readers can download the system, making the document a *Tutorial*, *Software Manual*, and *Resource*. Although a document could validly belong to multiple pedagogical roles,

we have carefully gone through several iterations of pedagogical roles to maximize the differences between roles. In other words, the distribution of the number of pedagogical roles per document is skewed such that most of the documents have one role. The *Other* role is an alternative category for all other possible pedagogical role types; we do not focus on documents with this role in this work. We believe most of the *Other* documents have high pedagogical value to a small group of experts and are beyond the scope of this initial investigation. In addition to these guidelines, we also provided a few examples of documents of each pedagogical role to annotators.

2.2 Annotation

The corpus of documents we annotated is drawn from a collection of pedagogically diverse documents related to natural language processing. The collection is based on the ACL Anthology, using the plain-text documents included in the ACL Anthology Network corpus (Radev et al., 2009). The ACL Anthology primarily consists of expert-level empirical research papers, so the collection was expanded to include other document types, as described in Gordon et al. (2017). Although we generally targeted specific document sources for specific pedagogical roles, we still found a variety of pedagogical roles from each source, i.e., not all documents from Wikipedia are *Reference Works*, and not all papers found while searching the web for "tutorials" are *Tutorials*. For annotation, we tried to identify a balanced sample of documents with different roles in this corpus by using simple regular expression pattern matching in document titles and abstracts. For example, to roughly target *Software Manuals*, we looked for documents with the phrase "software manual," "manual," or "technical manual" in the title or abstract.

To choose a reliable group of annotators, we internally annotated pedagogical roles for a set of documents and compared it with annotations done by a group of students pursuing master's degrees in computer science. We selected 11 students whose annotations had the highest correlation with our annotations. These annotators were instructed to read the abstract if there was one and to skim the rest of the document in enough detail such that they were able to annotate features for the document accurately and in a timely manner. We met regularly to discuss and come to a consensus on general document characteristics that were confusing to interpret.

We divided the documents for annotation into subsets of 100 to distribute among annotators so that each document was annotated by three annotators, and each subset was annotated by the same three annotators. We also manually filtered through and internally annotated 155 more supplementary documents to make up for a lack of documents that were annotated as *Surveys*, *Resources*, and *Software Manuals*. This supplementary set consists of 76 documents from the expanded ACL corpus and 79 additional documents collected from searching the web for more *Surveys*, *Resources*, and *Software Manuals*.[2]

2.3 Automatic prediction of pedagogical roles

We represent each document as a bag of sentence-embedding clusters. This technique embeds all sentences into vectors, clusters sentence vectors, and then represents documents as distributions over clusters. To evaluate the effectiveness of representing each document as a bag of sentence-embedding clusters and performing k-nearest neighbors classification, we also run two baseline techniques. One baseline technique is a multi-label centroid-based algorithm with sentence embeddings that is related to the single label centroid-based algorithm presented by Han and Karypis (2000) and the naïve Rocchio (1971) classification algorithm, a popular method for text classification (Rogati and Yang, 2002). The other baseline technique is a random forest classification of TF–IDF scores, which allows us to evaluate if sentence embeddings are more useful than word frequencies for this task.

We use sentence embeddings because specific sentences in documents are key indicators of the pedagogical roles of the document. As an explicit example, one might find the following in a *Survey* paper: "This paper presents a survey of the field of machine translation. . . " A more implicit example might be a *Resource* paper that mentions that one can find the corpus created by the authors at a specific link. We want to give much weight to the sentences that are the best indicators of the pedagogical roles of the document and leverage this information to automatically predict the pedagogical roles of documents. Skip-thought vectors[3] are able to effectively capture the semantics and syntax of sentences in several different tasks (Kiros et al.,

[2]Supplementary annotations are included in our publicly available annotation dataset.

[3]https://github.com/ryankiros/skip-thoughts

2015). To generate sentence embeddings needed for the centroid-based algorithm and the bag of sentence embedding clusters, we apply skip-thought vectors to embed each sentence from our annotated documents into a 4800-dimensional vector. We use the pre-trained skip-thought vector model to create sentence embeddings for each sentence.[4]

In our techniques, we do not pre-select sentences to include as features for classifying a document. We also do not treat sentences differently given their location in different sections of a document, e.g., introduction versus conclusion. Our corpus is composed of research papers, book chapters, Wikipedia articles, and web documents, so there is not a standard format that all documents follow. Our goal is to discover different types of sentences that could support our defined set of pedagogical roles as well as point to the existence of other roles.

Random Forest baseline classifier (RF): TF–IDF scores of words in our annotated documents are used as features for a random forest classifier. To calculate the TF–IDF scores, we included words that were in at least 10% and at most 90% of the documents. We used five-fold cross-validation to evaluate the results.[5]

Multi-label centroid-based algorithm with sentence embeddings (CEN): Each pedagogical role is represented by an average centroid vector, which is calculated by adding all sentence vectors in every document that belongs to the role, and then dividing the sum by the total number of sentence vectors added. When classifying a new document, we assign each sentence vector in the new document to a role label based on the nearest average vector. The role labels that are predicted for more than a third of the document's sentences are then predicted to be the document's role(s). Although this baseline method limits each document to two or fewer role predictions, it works as a rough baseline. 99.1% of the annotated documents have one or two pedagogical roles, and we assume our sample of annotated documents is representative of a larger collection of documents.

Bag of Sentence Embedding Clusters (BoSEC): Starting with the hypothesis that semantic and syntactic features of sentences are useful indicators of pedagogical roles, we employ k-means clustering[6] over sentence vectors to generate a representation basis (of N clusters) for computing a single $N \times 1$ feature vector per document. Each entry in the feature vector is the relative frequency of the specific sentence vector cluster being observed in the document.

K-Nearest Neighbors with Bag of Sentence Embedding Clusters (KNN+BoSEC): We use k-nearest neighbors classification to search for documents which exhibit the most similar distributions of clusters and predict the pedagogical roles of documents. To predict the roles of document A, we look for the three nearest documents in the N-dimensional vector space as calculated by the Manhattan distance metric. The majority roles of the three nearest documents are then predicted to be the roles of document A. The details of KNN+BoSEC are shown in Figure 1.

KNN+BoSEC with custom sentence encoder (KNN+BoSEC+): The content and style of writing in the scientific papers in our corpus differs from that of books used to train the pre-trained skip-thoughts vector model. We also run experiments using the KNN+BoSEC technique with a custom sentence embedding model trained on our entire collection of (annotated and unannotated) NLP documents. The custom sentence embedding model is trained using the default parameters described in the skip-thoughts training code.[7]

3 Results

3.1 Annotation agreement evaluation

The kappa value, which measures the likelihood of annotator agreement occurring above chance, is 0.68 for the pedagogical role annotations. This kappa value was calculated as an average over the kappa values for each subset of 100 documents. Given the difficulty of annotating pedagogical roles, which was confirmed by annotators, we believe a kappa of 0.68 indicates substantial agreement between annotators (Landis and Koch, 1977).

Table 1 shows the details of inter-annotator agreement for annotated pedagogical roles from documents with only one majority role. The rows are the majority roles, which we take to be the ground truth pedagogical roles of documents. The columns show the third annotator's annotations; if the third

[4]Model parameter details in Supplemental Material A.1.
[5]Model parameter details in Supplemental Material A.2.
[6]http://scikit-learn.org, model parameter details in Supplemental Material A.3.
[7]https://github.com/ryankiros/skip-thoughts; model parameter details in Supplemental Material A.4.

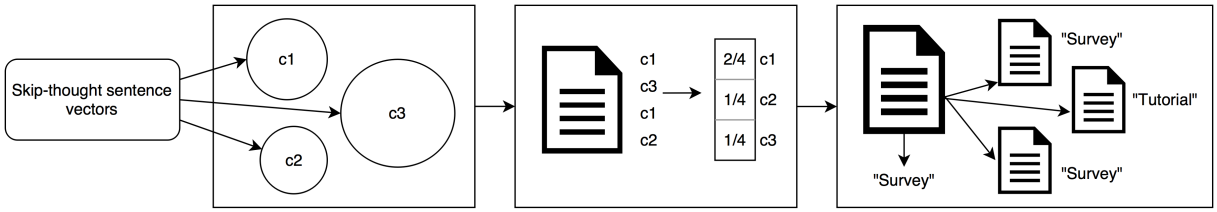

Figure 1: An end-to-end overview of the BoSEC+KNN technique. In (a), we generate skip-thought sentence vectors for every sentence in all documents. We partition all sentence vectors into clusters in (b). In (c), we represent each document as a distribution over the clusters formed in (b). (d) shows the KNN pedagogical role classification of documents based on the majority votes of annotated documents.

annotation matches the majority, then the particular annotation falls on the diagonal of Table 1. Although there are 1264 majority pedagogical role annotations, we calculated the confusion matrix for 1206 roles from documents with only one majority role each, for ease of interpretation. From the 1206 pedagogical roles, there are 1245 role pairs between the majority role and the third annotator's annotated role(s).

	Survey	*Tutorial*	*Resource*	*Reference Work*	*Empirical Results*	*Software Manual*	*Other*	Total
Sur.	**10**	1	0	7	4	0	5	27
Tut.	2	**44**	6	22	6	4	14	98
Res.	0	0	**5**	1	1	3	5	15
Ref.	36	20	3	**151**	4	1	28	243
Emp.	13	8	8	15	**526**	3	56	629
Sof.	0	1	0	0	2	**1**	2	6
Other	12	24	6	47	29	2	**107**	227
Total	73	98	28	243	572	14	217	**1245**

Table 1: Confusion matrix for annotated pedagogical roles from documents with only one majority role. Rows are the majority roles (chosen by two or three annotators) that we treat as ground truth. Columns are the third annotator's corresponding annotations.

From Table 1, we can see that *Survey* documents are sometimes confused with *Reference Works*, *Resource* papers are sometimes confused with *Other* documents, and *Software Manuals* are rare. We also see that *Other* documents have relatively higher rates of misclassification. These results are consistent with feedback from annotators. The reason why *Survey* documents are sometimes mistaken for *Reference Works* is because both examine a broad number of subjects in a domain; the distinction we make in our annotation guidelines is that *Reference Works* are a collection of established authoritative facts such as those one might find in an encyclopedia, whereas *Surveys* focus on the discoveries of other publications. When looking for *Resource* papers, annotators rely on looking for few indicator sentences that may be missed with a more superficial skim of the document. Also, the *Other* documents belong to a range of additional pedagogical roles, though we do not make finer distinctions here.

For each annotated document, we kept the pedagogical roles that had majority annotation agreement across the three annotators who annotated the document. If a document had no majority labels, the document was filtered out of the annotated document set. This filtered document set of 1235 documents with 1264 annotated pedagogical roles is the one we use along with a supplementary set for all pedagogical role prediction techniques.

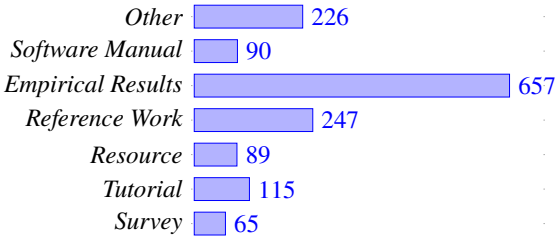

Figure 2: Distribution of all pedagogical role annotations in the full annotated corpus used for training classifiers

We noticed a lack of *Surveys*, *Resources*, and *Software Manuals*, so we internally annotated another supplementary set of 155 documents consisting mostly of documents of the underrepresented roles. The full annotated corpus we use for classification has the distribution of roles shown in Figure 2; this full corpus includes the filtered set of 1235 documents annotated by three annotators each and 155 internally annotated documents, for a total of 1489 pedagogical role annotations over 1390 documents. Given the corpora we selected our set of documents to annotate from, it is not surprising that most of the documents are *Empirical Results*, *Reference Works*, *Tutorials*, or *Other*. 94% of the annotated documents have just one pedagogical role, and 99.1% have one or two pedagogical roles.[8] The top three most common combinations of roles for a document are *Resource* and *Empirical Results*; *Resource* and *Software Manual*; *Tutorial*, *Resource*, and *Software Manual*.[9] Many documents with multiple pedagogical roles are *Resource* documents because the authors make their work publicly available.

3.2 Pedagogical role classification evaluation

In Table 2, we see that for both random forest classification of TF–IDF scores (RF) and sentence embedding methods (CEN, KNN+BoSEC, KNN+BoSEC+), the more samples there are for a pedagogical role, the higher the scores are for the role. The scores for *Other* documents are an anticipated exception to the trend, because we do not make more fine-grained distinctions between other pedagogical roles in this work. *Software Manuals* are also an exception to this trend, as their scores are relatively high for the number of samples; this is because *Software Manuals* are typically written in a very distinct style. CEN generally performs poorly across roles, doing worse than the baseline random forest classification with TF–IDF. This suggests that word frequency is more informative about the pedagogical roles of a document than a single representative vector per role.

With the exception of *Software Manuals*, RF is able to predict roles with more samples (*Reference Work*, *Empirical Results*, *Other*) with higher precision compared to roles with fewer samples (*Survey*, *Tutorial*, *Resource*). KNN+BoSEC and KNN+BoSEC+ have comparable precision for roles with more samples, but have significantly higher precision for roles with fewer samples. Compared to RF, KNN+BoSEC and KNN+BoSEC+ also have higher recall across all roles. KNN+BoSEC+ has the highest F_1 scores for all pedagogical roles. We attribute the fact that KNN+BoSEC+ is generally able to do better than KNN+BoSEC to using a custom sentence encoder trained on scientific documents.

Given that we use keyphrases to find documents that likely belong to specific pedagogical roles, we also want to see if we could achieve performance similar to that of our sentence embedding-based methods by simply classifying documents based on keyphrases. We manually curate a list of keyphrases for two pedagogical roles: "software manual," "manual," and "technical manual" for *Software Manuals*, and "tutorial" for *Tutorials*. We then classify a document as a certain role if any of the role's keyphrases are present in the first five sentences of the document, where the title counts as the first sentence. Classifying *Software Manuals* with this method has a precision of 0.15, a recall of 0.09, and an F_1 score of 0.11. KNN+BoSEC+ dramatically outperforms this method with the specified keyphrases for *Software manuals*. Classifying *Tutorials* with this method has a precision of 0.60, a recall of 0.50, and an F_1 score of 0.55. While the keyphrase classification results for *Tutorials* are closer to the corresponding KNN+BoSEC+ results, we think that the KNN+BoSEC+ results would also improve if it had access to the list of keyphrases as features, though we leave that for future experimentation. These initial keyphrase classification experiments suggest that sentence-embedding-based methods are generally more effective and robust than handcrafting keyphrases for each pedagogical role.

The confusion matrix in Table 3 allows us to make judgments about documents of different pedagogical roles, as predicted by KNN+BoSEC+. The rows are the ground truth roles, and the columns are the predicted roles. We can see that *Surveys*, *Resources*, and *Other* documents are often mistaken to be documents with *Empirical Results*. Additionally, there are relatively more instances of *Surveys*, *Resources*, and *Other* documents where the classifier is unable to make a prediction. Overall, these results suggest that the misclassifications are an effect of an unbalanced dataset with many more samples of *Empirical Results*, rather than an inherent lack of distinctness between documents of different roles.

Through a qualitative analysis of sentences from the clusters most frequently associated with each

[8] See Figure 3 in Supplemental Material for more details.
[9] See Figure 4 in Supplemental Material for more details.

Ped. role	Precision				Recall				F_1				Support
	RF	CEN	KNN+BoSEC	KNN+BoSEC+	RF	CEN	KNN+BoSEC	KNN+BoSEC+	RF	CEN	KNN+BoSEC	KNN+BoSEC+	All
Survey	0	0.02	0.23	**0.31**	0	**0.21**	0.20	0.18	0	0.03	0.21	**0.23**	13
Tutorial	0.50	0.10	0.64	**0.66**	0.05	0.21	**0.55**	0.52	0.08	0.11	0.57	**0.58**	23
Resource	0.20	0	**0.70**	0.53	0.01	0	0.19	**0.24**	0.03	0	0.29	**0.32**	17.8
Ref. Work	0.77	0.07	0.71	**0.78**	0.33	0.32	0.70	**0.71**	0.46	0.11	0.70	**0.74**	49.4
Emp. Res.	**0.86**	0	0.83	0.85	0.77	0	0.86	**0.89**	0.81	0	0.85	**0.87**	131.4
Sof. Man.	**0.98**	0.05	0.93	0.95	0.34	0.16	0.72	**0.86**	0.49	0.07	0.81	**0.90**	18
Other	0.63	0.06	0.57	**0.65**	0.10	0.40	0.27	**0.48**	0.17	0.10	0.36	**0.55**	45.2
avg / total	0.71	0.03	0.73	**0.76**	0.44	0.15	0.64	**0.70**	0.50	0.05	0.66	**0.72**	297.8

Table 2: Precision, recall, and F_1 scores by pedagogical roles for all methods. Support is the actual number of documents with each role. avg / total computes weighted averages of scores across all roles. All values are averaged over a five-fold cross validation.

	Survey	*Tutorial*	*Resource*	*Reference Work*	*Empirical Results*	*Software Manual*	*Other*	No prediction	Total
Sur.	**2**	0.2	0	0.8	4.2	0	1.2	1.6	10
Tut.	0.2	**9.4**	0	2.6	0.8	0.8	2	2.2	18
Res.	0.2	0	**1.2**	0	3	0	0.6	1.8	6.8
Ref.	1.2	1.6	0.2	**34**	3.6	0.2	3.2	4.2	48.2
Emp.	0.8	1.4	1.8	3	**109.2**	0	4	4.2	124.4
Sof.	0	1.6	0.6	0.2	0	**9.8**	0	0.4	12.6
Oth.	3.4	0.6	0.6	3.2	8	0	**21.8**	7.6	45.2
Tot.	7.8	14.8	4.4	43.8	128.8	10.8	32.8	22	265.2

Table 3: Ground truth pedagogical roles (rows) versus predicted roles (columns) using KNN+BoSEC+. We calculate the confusion matrix for documents with only one ground truth role. All values are averaged over a five-fold cross validation.

pedagogical role, we observe that example sentences from different roles align with our intuitions of what exemplary sentences from different roles should be. The *Survey* sentences describe progress in different areas of research; the *Tutorial* sentences explain details of specific concepts and methods; the *Software Manual* sentences give information about how to use a tool.[10] Sentences from the most frequent clusters of a role do not explicitly mention the roles of the paper, e.g., "This paper presents a tutorial. . . " This phenomenon makes sense for two reasons. One reason is that the majority of documents do not explicitly say what kind of document they are. The second reason is that even when documents do explicitly state their role, the actual content of the document may disagree with the declared role. For example, some papers are written to accompany tutorials presented at workshops. The papers will explicitly declare themselves to be tutorials, but the paper will only include an abstract and not the tutorial itself. Following our annotation guidelines, we do not label these documents as *Tutorials*. This implicit characterization of a document's pedagogical roles through sentences means that a method that merely searches for explicit mentions of keywords or declaration of the document's roles would not be an effective approach to this problem. Thus, these example sentences qualitatively validate our embedding and clustering approach to pedagogical role classification.

4 Related Work

To the best of our knowledge, there is not much prior work that is directly related to investigating relevant pedagogical features of documents through pedagogical roles. There are some document recommendation systems that try to find documents that

[10]For more details, see Table 4 in Supplemental Material.

are both conceptually relevant to a user's query and pertinent to the user's interest, level of background knowledge, etc. For example, Semantic Scholar[11] allows users to filter an automatically generated reading list by "overviews," which are analogous to our definition of Surveys. PageRank accounts for popularity when identifying documents of interest (Page et al., 1999). Tang and McCalla (2004) consider the user's background knowledge, interest towards specific topics, and motivation when making recommendations. Gori and Pucci (2006) present a research paper recommender system based on the random walk algorithm and a small set of papers that users mark as relevant. Santos and Boticario (2010) emphasize that recommendation systems in the e-learning domain should be "guided by educational objectives" and define a semantic model for recommendation objects.

Previous efforts at investigating the value of documents include evaluating the reading difficulty of documents, citation graphs, and surveys, though none really address the problem of estimating the pedagogical value of a document to a learner while focusing on the interpretability of the results. The interpretability of results is especially important in education because educators need to be able to provide clear feedback to students. In automatic essay scoring, researchers look at features such as word count, semantic and syntactic coherence, sentence length, vocabulary complexity, and the use of certain phrases that facilitate the flow of ideas, e.g., "first of all" (Burstein et al., 2004; Shermis and Burstein, 2013). These features are a starting point to estimate the value of a document, but to estimate pedagogical value, we must consider if and how these features would affect different learners. Other directions of research use the influence of a paper within a citation graph as a proxy for the value of the paper, following the reasoning that good quality papers would be more important "nodes" in a citation graph (Ekstrand et al., 2010); however, documents that are important "nodes" in the graph do not necessarily have high pedagogical value for all learners. Tang and McCalla (2009) present surveys to students as an annotation method to estimate the value of the paper to the learner. They annotate individual features of job-relatedness, interestingness, usefulness, etc., using ordinal-scale values, and study the partial correlations between features to analyze the composition of features that contribute to the pedagogical value of a document. Our approach is different in that (a) we develop an intermediate representation of pedagogical value that can be largely objectively annotated, (b) we evaluate correlation between annotators and not between features, and (c) we additionally present baseline results of pedagogical role prediction.

The classification task described in this work is also related to text classification, a task with a long history in NLP. Sebastiani (2002) presents a detailed survey of tasks and techniques used in text classification up until the early 2000s. Joachims (1998) presents experimental results that justify the use of Support Vector Machines (SVMs) for text classification. Soucy and Mineau (2001) use TF–IDF scores and a KNN model to perform different text categorization tasks.

5 Conclusion

In this paper, we have described (a) our creation of the first annotated corpus of pedagogical roles for the study of pedagogical value and (b) our use of sentence embeddings and clustering techniques to develop a baseline for pedagogical role classification. The inter-annotator agreement for the annotation of pedagogical roles is substantial and thus a good basis to develop pedagogical role classification techniques and intuitions about pedagogical value upon. Analyses of our bag of sentence-embedding clusters technique support our intuition that certain sentences in a document are strong indicators of the pedagogical roles of the document. The next steps are to expand the set of roles as needed and apply our techniques to other domains in order to work towards a general approach to estimating pedagogical value. We believe it is important to make our corpus and annotations public, as feedback from other researchers will help improve the quality and scope of our corpus as we expand it.

Acknowledgments

The authors thank Yigal Arens, Aram Galstyan, and Linhong Zhu for their valuable feedback on this work.

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Research Laboratory (AFRL). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the

[11] https://www.semanticscholar.org

official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

John B. Biggs. 2011. *Teaching for quality learning at university: What the student does*. McGraw-Hill Education (UK).

Kurt D. Bollacker, Steve Lawrence, and C. Lee Giles. 1999. A system for automatic personalized tracking of scientific literature on the web. In *Proceedings of the Fourth ACM Conference on Digital Libraries*, DL '99, pages 105–113, New York, NY, USA. ACM.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The criterion online writing service. *AI Magazine*, 25(3):27.

Michael D. Ekstrand, Praveen Kannan, James A. Stemper, John T. Butler, Joseph A. Konstan, and John T. Riedl. 2010. Automatically building research reading lists. In *Proceedings of the Fourth ACM Conference on Recommender systems*, pages 159–166. ACM.

Jonathan Gordon, Stephen Aguilar, Emily Sheng, and Gully Burns. 2017. Structured generation of technical reading lists. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*.

Jonathan Gordon, Linhong Zhu, Aram Galstyan, Prem Natarajan, and Gully Burns. 2016. Modeling concept dependencies in a scientific corpus. In *Proceedings of ACL*.

Marco Gori and Augusto Pucci. 2006. Research paper recommender systems: A random-walk based approach. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, pages 778–781. IEEE.

Eui-Hong Sam Han and George Karypis. 2000. Centroid-based document classification: Analysis and experimental results. In *European conference on principles of data mining and knowledge discovery*, pages 424–431. Springer.

James G. Jardine. 2014. Automatically generating reading lists. Technical Report UCAM-CL-TR-848, University of Cambridge Computer Laboratory.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pages 137–142.

Ryan Kiros, Yukun Zhu, Ruslan R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66, Stanford InfoLab.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL Anthology Network Corpus. In *Proceedings, ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, Singapore.

Joseph J. Rocchio. 1971. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall, Englewood Cliffs, NJ.

Monica Rogati and Yiming Yang. 2002. High-performing feature selection for text classification. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 659–661. ACM.

Olga C. Santos and Jesus G. Boticario. 2010. Modeling recommendations for the educational domain. *Procedia Computer Science*, 1(2):2793–2800.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47.

Mark D. Shermis and Jill Burstein. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Pascal Soucy and Guy W. Mineau. 2001. A simple KNN algorithm for text categorization. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 647–8, Washington, DC, USA. IEEE Computer Society.

Tiffany Y. Tang and Gordon McCalla. 2009. The pedagogical value of papers: a collaborative-filtering based paper recommender. *Journal of Digital Information*, 10(2).

Tiffany Ya Tang and Gordon I. McCalla. 2004. On the pedagogically guided paper recommendation for an evolving web-based learning system. In *FLAIRS Conference*, pages 86–92.

A Supplemental Material

A.1 Skip-thought vector parameters

Each sentence vector has 4800 dimensions, with the first 2400 dimensions as the uni-skip model,

and the latter 2400 dimensions as the bi-skip model. The model has the following parameters: recurrent matrices initialized with orthogonal initialization, non-recurrent matrices initialized from a uniform distribution in $[-0.1, 0.1]$, mini-batches of size 128, gradients clipped when the norm of the parameter vector is greater than 10, and the Adam algorithm for optimization.

A.2 Random forest classification parameters

For the random forest classifier, we used the Gini impurity function to estimate the quality of splits. When looking for the best split, the classifier considers the square root of the total number of features. The maximum depth of the tree is 75, and the classifier splits on a minimum of 5 samples at the internal nodes. We use 10 trees and a minimum of 1 sample at each leaf node.

A.3 Mini-batch K-means parameters

In this clustering technique, random subsets of the feature vectors are used in each iteration. We train the model with 300 clusters, early stopping if there is no improvement in the last 50 mini batches, a mini batch size of 4800, and the fraction of the maximum number of counts for a cluster center to be reassigned is 0.0001. We had experimented with different cluster sizes, and found 300 clusters to be the right size to maintain coherency within and distinction across clusters.

A.4 Custom skip-thought vector model parameters

Specifically, the RNN word embeddings have 620 dimensions, and we use a uni-skip model with a hidden state size of 2400. Both the encoder and the decoder are GRUs. The size of the decoder vocabulary is 20000, and the maximum length of a sentence is 30 words; additional words in sentences are ignored. Our custom model is trained for 5 epochs, has a gradient clipping value of 5, has a batch size of 64, and uses the Adam optimization algorithm.

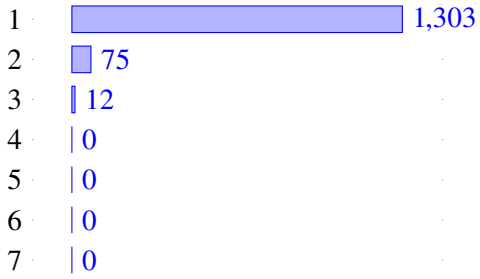

Figure 3: Distribution of number of pedagogical roles per document in full annotated corpus

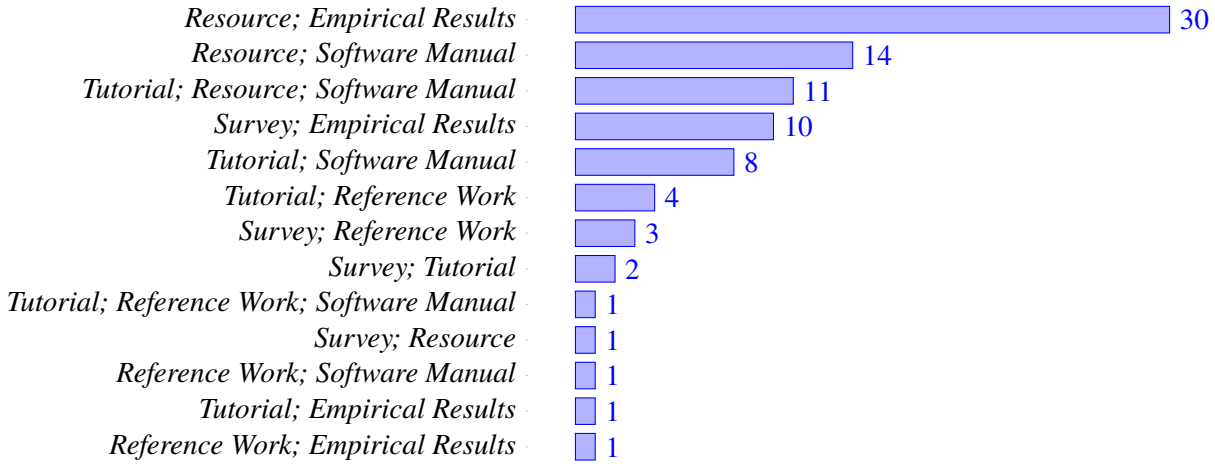

Figure 4: Distribution of pedagogical roles for documents in full annotated corpus with more than one role

Pedagogical role	Cluster ID	Example sentence
Survey	250	This view has been worked out in the text generation and dialog community more than in the text understanding community (Mann and Thompson, 1987; Hovy, 1993; Moore, 1994).
	123	Confronted with the claim that Game Theory should be the theoretical backbone to NLG, some people might respond that no new backbone is needed, because the theory of formal languages, conjoined with a properly expressive variant of Symbolic Logic, provides sufficient backbone already.
Tutorial	209	As you guessed from my explanations of different notations, different regex engine designers unfortunately have different ideas about the syntax to use.
	95	This information is incorporated in the tri-factorization model via a squared loss term, where the notation Tr (4) means trace of the matrix A.
Resource	147	`>>> windowdiff(s1, s1, 3)`
	255	... `print('', repr(corpus.fileids())[:60])`
Reference Work	155	The greater the resumption of the activity (i.e., mismatch negativity), the more different the neurological processing of the new item.
	86	A trajectory of an object is determined by its different centers of gravity relative to an underlying coordinate system.
Empirical Results	183	5.3 Using Multiple Knowledge Sources
	62	The NCC open track is shown in the following table 2.
Software Manual	147	`>>> clf.fit(X, Y)`
	152	An example of this approach can be found in the /verbi folder in the Italian MOR grammar.
Other	279	The problem in the cases (3) and (4) is how and why the hearer fails to derive implicatures.
	157	Proofs of the form suppose-absurd F D are called proofs by contradiction.

Table 4: Example sentences from the clusters most frequently associated with each pedagogical role. The clusters representing mostly punctuation, numbers, or incoherent strings were not included in calculating most frequently associated clusters.