YouEDU: Addressing Confusion in MOOC Discussion Forums by Recommending Instructional Video Clips

Akshay Agrawal
Stanford University
akshayka@cs.stanford.edu

Jagadish Venkatraman
Stanford University
vjagadish@cs.stanford.edu

Shane Leonard
Stanford University
shanel@stanford.edu

Andreas Paepcke
Stanford University
paepcke@cs.stanford.edu

ABSTRACT

In Massive Open Online Courses (MOOCs), struggling learners often seek help by posting questions in discussion forums. Unfortunately, given the large volume of discussion in MOOCs, instructors may overlook these learners' posts, detrimentally impacting the learning process and exacerbating attrition. In this paper, we present YouEDU, an instructional aid that automatically detects and addresses confusion in forum posts. Leveraging our publicly-available Stanford MOOCPosts corpus, we train a heterogeneous set of classifiers to classify forum posts across multiple dimensions. In particular, classifiers that target sentiment, urgency, and other descriptive variables inform a single classifier that detects confusion. We then employ information retrieval techniques to map confused posts to minute-resolution clips from course videos; the ranking over these clips accounts for both video-clickstream data and textual similarity between posts and closed captions. We measure the performance of our classification model in multiple educational contexts, exploring the nature of confusion within each; we also evaluate the relevancy of materials returned by our ranking algorithm.

1. INTRODUCTION

During recent years a number of universities have experimented with online delivery of their courses to the public. Hundreds of thousands of learners across the world have taken advantage of these opportunities. While teaching techniques and technologies for such large numbers will change in the coming years, fundamental challenges will remain.

For example, learners can get lost in a sense of isolation, as no physically accessible peer group accompanies them through the sometimes difficult material. Vast diversity in prior studying experience disadvantages learners who are unused to systematic, self driven work. Instructors, on the other side of the Internet, are unable to interact with individual struggling learners as they might in traditional settings.

Free, or very inexpensive online learning opportunities carry the potential of enormous public good. But unless these impediments are addressed, the significant effort and expense invested in the course offerings will not impart their maximum impact.

Two technologies have so far been common to most online teaching: instructional videos, and a course-internal communication forum that allows learners to interact with each other. Both of these technologies exhibit both strengths and weaknesses.

Videos, while old fashioned and maybe not optimal as teaching tools, do communicate archivable material, and many such assets are available at this point. Yet one of the technology's downsides is that video is tyrannically linear in nature. No table of contents or hyperlinks are available to access material randomly. With often over a hundred ten to fifteen minute videos in a course, learners can easily be discouraged when they find a hole in their understanding, and need to re-view relevant footage.

Course forum facilities can be powerful reflections of learner mood and success. In the best cases learners answer each others' questions, furthering a sense of belonging. Instructors could in theory read the forum posts, and gather a good sense of what is going well, and where learners are struggling. But with 30 thousand posts in one course alone, instructors need help.

Our work aims to solve both the described problems in a unified approach. We automatically classify forum posts, both in service of identifying posts that instructors do need to act on, and to recommend relevant sections of course videos. These recommendations are computed by using subsets of post contents as queries into closed caption files. Such files are transcriptions of speech in the videos, and are created for compliance with US legislation that ensures equal access to learners with disabilities, such as sight impairments.

In particular, we classify the forum posts along several dimensions. One classifier attempts to identify posts that are evidence of a learner's *confusion* about some learning goal. Other classifiers determine whether a post was a *question*,

an *answer*, or an *opinion*. Additional classifiers try to determine each post's *sentiment*, and *urgency* for an instructor to respond to the post.

These classifications can be put to several uses. For example, posts with high urgency can be deployed as filters that allow instructors to respond to the most important posts. The *answer* class can be used to identify learners of possibly high proficiency. Such learners can then be encouraged, or directly asked to address some posted questions. In this work we focus on the *confusion* classifier, and link it to the problem of finding post-relevant video snippets at one minute resolution. That is we identify confused posts, and produce a ranked list of video snippets that are likely to help address the confusion.

The fine grained time resolution of recommendations is important for several reasons. First, the obvious time savings such resolution affords can encourage learner persistence. Second, studies have shown that short video segments are more effective for learners, than extended footage [**?**].

While classification technology is well known, two factors have prevented our approach up to now. The type of classifiers we require are supervised, and therefore need a training set. Second, the feature set we employ to advantage include data beyond the forum posts themselves.

Such data has only recently become available. In order to obtain the training set for our classifiers we hired paid consultants to tag 30,000 posts from three categories of large, public Stanford online courses: Education, Engineering, and Medicine. The set is available to researchers on request [**?**].

In addition to the forum posts, all the online teaching platforms Stanford uses to distribute their public courses gather tracking log data that now include hundreds of millions of learner actions, including homework assignment outcomes. We use a subset of these data as features for our confusion classification. Some of these data are also available in anonymized form to researchers upon request [**?**]. Until very recently neither data set has been available, preventing the solution approach we chose for this work.

This paper explores a number of implementation options for the forum classification, and evaluates the effectiveness of the caption file text retrieval.

The remainder of this paper is organized as follows. We examine related work, present the Stanford MOOCPosts corpus in Section 3, and sketch the architecture of YouEDU in Section 4. In Sections 5 and 7 we detail and evaluate YouEDU's constituent classification and recommendation phases. We close with a section on future work, and a conclusion.

2. RELATED WORK

... Akshay, please include something like:

Closed caption files were used in the Informedia project [**?**] to index into television news shows.

3. THE STANFORD MOOCPOSTS CORPUS

Given that no publicly-available corpus of tagged MOOC discussion forum posts existed prior to our research, we set out to create our own. The outcome of our data compilation and curation was the Stanford MOOCPosts dataset: a corpus composed of 29,604 anonymized learner forum posts from eleven Stanford University public online classes. Freely available to academic researchers, the MOOCPosts dataset was designed to enable computational inquiries into the nature of both affect and content in MOOC discussion forums.

Each post in the MOOCPosts dataset was scored across six dimensions – confusion, sentiment, urgency, question, answer, and opinion – and subsequently augmented with additional metadata.

3.1 Methodology: Compiling the Dataset

We organized the posts into three sets of related courses: Humanities/Sciences, Medicine, and Education, with 10,000, 10,002, and 10,000 entries, respectively. Humanities/Sciences contains two economics courses, two statistics courses, a global health course, and an environmental physiology course; Medicine contains two iterations of the same medical statistics course, a science writing course, and an emergency medicine course; and Education contains a single course titled *How to Learn Math*.

Each course set was coded by three distinct, independent, paid oDesk coders. That is, three triplets of coders each worked on one set of 10,000 posts. No coder worked on more than one course set. Each coder attempted to code every post for his or her particular set. All posts with malformed or missing scores in at least one coder's spreadsheet were discarded. This elision accounts for the difference between the 29,604 posts in the final set, and the original 30,002 posts.

Coders were asked to score their posts across six dimensions:

- Question: Does this post include a question?
- Opinion: Does this post include an opinion, or is its subject matter wholly factual?
- Answer: Does this post appear to be an answer to a learner's question?
- Sentiment: What sentiment does this post convey, on a scale of 1 (extremely negative) to 7 (extremely positive)? A score of 4 indicates neutrality.
- Urgency: How urgent is it that an instructor respond to the post, on a scale of 1 (not urgent at all) to 7 (extremely urgent)? A score of 4 indicates that the intsructor should respond only if he or she has spare time.
- Confusion: To what extent does this post express confusion, or the lack thereof, on a scale of 1 (expert knowledge) to 7 (extreme confusion)? A score of 4 indicates neutrality in that the post expresses neither knowledge nor confusion.

Coders were given examples of posts in each category. The following was an example of an extremely urgent post:

> The website is down at the moment https: //class.stanford.edu/courses/Engineering/ Networking/Winter2014/courseware seems down

	Humanities	Medicine	Education
Urgency	0.657	0.485	0.000*
Sentiment	-0.171	-0.098	-0.134
Opinion	-0.193	-0.097	-0.297
Answer	-0.257	-0.394	-0.106
Question	0.623	0.459	0.347

Table 1: Correlations with Confusion. The urgency and question variables are strongly correlated with confusion. All correlations, save the one denoted by *, were significant, with p-values < 0.01.

> and I'm not able to submit the Midterm. Still have the "Final Submit" button on the page, but it doesn't work. Are the servers congested? thanks anyway

And

> Double colons "::" expand to longest possible 0's If the longest is 0, will the address be considered valid ? (even if it doesn't make sense and there is no room for adding 0's) Can someone please answer ? Thanks in advance

was given as an example of a post that was both confused (6.0) and urgent (5.0).

We created three gold sets from the coders' scores, on for each course set. We computed inter-rater reliability using Krippendorff's Alpha [**?**]. For a given post and Likert variable, the post's gold score was computed as an unweighted average of the scores assigned to it by the subset of two coders who expressed the most agreement on that particular variable. Gold scores for binary variables were chosen by majority votes across all three coders. We refer readers to our write-up in [**?**] for a more detailed treatment of our procedure and our complete inter-rater reliability results.

3.2 Discussion

We found significant correlations between confusion and the other five variables. In the humanities and medicine course sets, confusion and urgency were correlated with a pearson's coefficient of 0.657 and 0.485, respectively. In all three subdivisions of the dataset, confusion and the question variable were positively correlated (0.623, 0.459, and 0.347), while the sentiment, opinion, and answer variables were negatively correlated with confusion. Table 1 reports the entire set of correlations.

That questions and confusion were positively correlated supports the observation in [**?**] that confusion is often communicated through questions. The negative correlations make intuitive sense, too. A confused learner might become frustrated and express negative sentiment; as discussed in [**?**], confusion and frustration sometimes go hand-in-hand. If a learner is opining on something, then it seems less likely that he or she is discussing a factual topic. And we would hope that learners providing answers to questions are not themselves confused.

4. YOUEDU: DETECT AND RECOMMEND

YouEDU is a personalized intervention system that recommends educational video clips to learners. Figure 1 illustrates the key steps that comprise YouEDU. YouEDU takes as input a set P of forum posts, processing them in two distinct phases: (I) detection and (II) recommendation. In the first phase, we apply a classifier to each post in P, outputting a subset P_c consisting of posts in which the classifier detected confusion. The confusion classifier functions as a *combination* classifier in that it combines the predictions from classifiers trained to predict other post-related qualities.

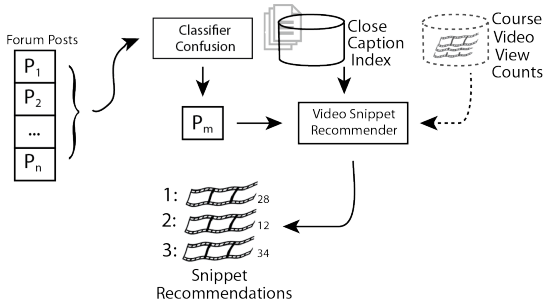

Figure 1: YouEDU Architecture. The YouEDU pipeline consists of two phases: post classification and video snippet recommendation. The dotted-line module is under construction (see Section 8)

The second phase takes P_c as input and, for each confused post in $p \in P_c$, outputs a ranked list of educational video snippets that address the object of confusion expressed in p. In particular, for a given post, the recommender produces an initial ranking across a number of one-minute video clips by computing a similarity metric between the post and closed caption sections. The ranking of videos in the retrieved set is then further informed by video-clickstream data.

While YouEDU outputs minute-resolution video clips, it does not necessarily guarantee that these clips fully address the exhibited confusion – indeed, several minutes of instructional content are often required to explain a single concept. Rather, the video snippets collectively form an ad-hoc index. For example, say that for a given post, YouEDU outputs three video snippets with start times s_1, s_2, s_3, in order of decreasing relevance, and say that these snippets were contained in videos v_1, v_2, v_3, respectively, v_1, v_2, v_3 not necessarily unique. In order to clarify his or her confusion, the author of the post should begin watching video v_1 at s_1 – the learner can autonomously set the end time of the snippet, and can move on to the next video, start time pair if any confusion still lingers.

In the following two sections, we delve further into both phases of YouEDU, describing them in detail and relating the results of empirical evaluations.

5. PHASE I: DETECTING CONFUSION

We frame the problem of detecting confusion as a binary one: Given a discussion forum post p with a true label L in {not confused, confused}, apply some hypothesis h that correctly divines L. Posts with a confusion rating greater than four in the MOOCPosts dataset fall into the "confused" class, while

all other posts fall into the "not confused" class.

We craft a rich feature space that fully utilizes the data available in our MOOCPosts dataset, choosing logistic regression with l_2 regularization as our statistical model. Results from empirical evaluations demonstrate that our classifier performs reasonably well, while simultaneously providing insight into the nature of confusion across multiple courses.

5.1 Feature Space and Model Design

Our feature space is composed of three types of inputs, those derived from: the post body; post metadata; and other classifiers. The confusion classifer we train functions as a combining layer that folds in the predictions of other classifiers; these classifiers are trained to predict variables correlated with confusion. We expand upon each type of input here.

5.1.1 Bag-of-Words

We take the bag-of-words approach in representing documents, or forum posts. Each document is represented in part as a vector of indicator variables, one for each word that appears in the training data – the i-th indicator is one if the i-th word in the vocabulary is present in the document, zero otherwise. A word is defined as either a sequence of one or more alphanumeric characters or a single punctuation character (one of {. , ; ! ?}).

Documents are pre-processed before they are mapped to vectors. We prune out stop words, using a subset of the stop word list published by the Information Retrieval Group at the University of Glasgow [**?**]. Removed words inlcude, but are not limited to, interrogatives ("who", "what", "where", "when", "why", "how"), words that identify the self ("I", "my"), verbs indicating ability or the lack thereof, negative words ("cant", "cannot", "couldnt"), and certain conjunctions ("yet", "but"). We ignore alphabetic case[1] and lemmatize numbers, LaTeX equations, and URLs. Intuitively, the presence of numbers and equations in a forum post might indirectly convey confusion or the lack thereof, in that the learner may be asking a question about some quantity or perhaps providing an answer to a quantitative question; similarly, a knowledgeable learner might answer a question by citing a URL.

The unigram document representation, while simple, pervades text classification and often achieves high performance [**?**]. We employ l_2 regularization in order to prevent overfitting, a risk that is aggravated when the dimension of the feature space exceeds the training set size [**?**].

5.1.2 Post Metadata

The feature vector derived from unigrams is augmented with post metadata, including:

- The number of up-votes accumulated by the post. We rationalized that learners might express interest in posts that voiced confusion that they shared.
- The number of reads garnered by the post's containing thread.
- Whether the poster elected to appear anonymous to his or her peers or to the entire population. It has been shown that anonymity in educational discussion forums enables learners to ask questions without fear of judgement [**?**], and our dataset demonstrates a strong correlation between questions and confusion.
- The post author's grade in the class at the time of post submission, where "grade" is defined as the numer of points earned by the learner (e.g., by correctly answering quiz questions) divided by the number of points possible. The lower the grade, we hypothesized, the more likely the learner might be confused about a topic.
- The post position – that is, whether or not the post was the first message in a thread. In order to seek help on a forum, a learner must first a post; most likely, we hypothesized, the learner will create a new thread for that post.

5.1.3 Classifier Combination

In section 3.2, we demonstrated that, at least in the humanities and medicine courses, confusion is significantly correlated with questions, answers, urgency, sentiment and opinion. As such, in predicting confusion, we take into account the predictions of five distinct classifiers, one for each of the aforementioned variables. We use the fine-grained method of combining classifiers in which the outputs of several classifiers are fed as input to a *combination function* [**?**]. In our case, the combination function is itself a classifier.

For a given train-test partition, let D_{train} be the training set and D_{test} be the test set. In both sets, each example is tagged along the six variables. Let H_q, H_a, H_o, H_s, and H_u be classifiers for the question, answer, opinion, sentiment, and urgency variables, respectively. We call these classifiers *constituent* classifiers. Each constituent is trained on D_{train}, taking as input bag-of-words and post metadata features, as described in the previous two subsections.

Let H_c, a classifier for confusion, be our combination function. Like the constituent classifiers, H_c is trained on D_{train} and takes as input bag-of-words and metadata features. Unlike the constituents, H_c also treats the ground-truth labels for the question, answer, opinion, sentiment, and urgency variables as features. When testing H_c on an example $d \in D_{test}$, the constituent classifiers each output a prediction for d. These five predictions are appended to the bag-of-words and metadata features derived from d. The resulting vector is given as input to H_c, which then predicts d's confusion class.

A few subtleties: H_s uses an additional metadata feature that the other classifiers do not – the number of negative words (e.g., "not", "cannot", "never", etc.). H_q, H_a, H_u, and H_c treat the number of question marks as an additional feature, given the previously presented correlations; [**?**] also used question marks in predicting confusion. And while H_q, H_a, and H_o are by nature binary classifiers, H_s and H_u are multi-class. They predict values corresponding to negative (raw score < 4), neutral (raw score $= 4$), and positive (raw

[1] All-caps discussion certainly does communicate affect in some Internet forums – it is typically associated with aggression and is considered a breach of "nettiquette" [**?**]; however, we assume that MOOC forum-goers are somewhat civil, and so accounting for case would needlessly inflate our feature space.

score > 4), in order to provide H_c with somewhat granular information.

5.2 Evaluation and Discussion

For clarity, we refer to the confusion classifier that uses all the features described in the section 5.1 as the *combined* classifier. In this section, we evaluate and interpret the performance of both the combined classifier and confusion classifiers with pared-down feature sets, reporting insights and intuitions gleaned about the nature of confusion in MOOCs along the way.

We quantify performance primarily using two metrics: F_1 and Cohen's Kappa. We favor the Kappa over accuracy because the former accounts for chance agreement [**?**]. Unless stated otherwise, reported metrics represent an average over 10 folds of stratified cross-validation.

5.2.1 Confusion at the Course-Set Granularity

Table 2 presents the performance of the combined classifier on the humanities and medicine course sets. As mentioned in section 3.1, both sets are somewhat heterogenous collections of courses, with a total nearly 10,000 posts in each set. In our dataset, not-confused posts outnumber confused ones – in the humanities course set, but 23% of posts exhibit confusion, and in the medicine course set, 16% do. As such, our classifier was naturally better at identifying non-confused posts than confused posts.

5.2.2 The Language of Confusion Across Courses

Table 3 presents the performance of the combined classifier on select courses, sorted in descending order by Kappa. Our classifier performed best on courses that traded in highly technical language. Take, for example, the following two posts from *Managing Emergencies*:

> What could have caused the hemoptysis in this patient with pnuemonia?

> At what doses is it therapuetic for such a patient because at high doses it causes vasoconstrition through alpha1 interactions, while at low doses it causes dilation of renal veins and splachinic vessels.

Both of these posts were tagged as exhibiting confusion, and both of them are fairly saturated with medical terms. Incidentally, our classifier achieved its highest performance when cross-validating on this course (Kappa = 0.741). This makes intuitive sense – a vocabulary so technical and esoteric is likely only used when a learner is discussing or asking a question about a specific course topic. Indeed, inspecting our model's weights revealed that "systematic" was the 11^{th} most indicative feature for confusion (odds ratio = 1.23) and "defibrillation" was the 15^{th} (odds ratio = 1.22). Similarly, in *Statistical Learning*, "solutions" was the sixth most indicative feature for confusion (odds ratio = 1.75), and"predict" was the ninth (odds ratio = 1.65).

A glance at Table 3 suggests that our classifier's performance degrades as the discourse becomes less technical. Posts like the following were typical in *How to Learn Math*, an education course about the pedagogy of mathematics:

> I am not sure if I agree with tracking or not. I like teaching children at all levels. It seems that if you teach kids at the same level then it becomes the Sam ethnic *[sic]* over and over. ... In a normal class setting the lower level learners can learn from the higher learners and vice versa. Although I do find it very hard to find a middle ground. There has to be an easier way.

The above post was tagged as exhibiting confusion, but the language is much more subtle than that seen in the posts from *Managing Emergencies*. In technical courses, there was a sharp shift in diction between confused posts and not-confused posts. The same could not necessarily be said for non-technical courses, and it is not surprising that we saw our lowest Kappa (0.359) when classifying *How to Learn Math*. Indeed, in the education course, learners tend to voice more confusion about the structure of the class than the content itself – "link", "videos", and "responses" are the fourth, fifth, and seventh most indicative features for the confusion class, respectively.

Examining the feature weights learned when cross-validating on the humanities and medicine course sets provides us with a more holistic view onto the language of confusion. Domain-specific words take the backseat to words that convey the learning process. For example, in both course sets, "confused" was the word with the highest feature weight (odds ratios equal to 3.19 and 2.97 for humanities and medicine, respectively). In the humanities course set, "?", "couldn't", "question", "haven't", and "wondering" came next, in that order. The importance of question-related features in particular is consistent with [**?**] and with the correlations in the MOOCPosts dataset. In medicine, the next highest ranked words were "explain", "role", "understand", "stuck", and "struggling".

Table 4 displays the most informative features for humanities and medicine course sets, as well as *How to Learn Math* and *Managing Emergencies*.

5.2.3 Training and Testing on Distinct Courses

We ran a series of experiments in which we trained the combined classifier on posts from one course and then tested it on posts from another one, without cross-validation. The results of these experiments are tabulated in Table 5.

Our highest Kappa (0.629) was achieved when training on *Statistics in Medicine 2013* and testing on *Statistics in Medicine 2014*; this makes sense, since they comprise two runs of the same course. Many instructors plan to offer the same MOOC multiple times [**?**]. If an instructor could tag but one of those runs, then a classifier like ours deployed in an online setting might be met with success. Even if that tagging process is too expensive, our results when training on *Statistical Learning* and testing on *Statistics 216* suggest that an online classifier might perform well so long as its training data derives from the same domain as the test data. Performance might suffer, however, if the domains of the training and test data are non-overlapping, as is the case in the last two experiments in Table 5.

5.2.4 Constituent Classifiers and Post Metadata

Course Set	Not Confused			Confused			Kappa
	Precision	Recall	F_1	Precision	Recall	F_1	
Humanities	0.898	0.943	0.919	0.778	0.642	0.700	0.621
Medicine	0.924	0.946	0.935	0.699	0.589	0.627	0.564

Table 2: Combined Confusion Classifier Performance, Course Sets. For each course set listed, this table reports the average performance across 10 folds of stratified cross-validation.

Course	# Posts (% Confused)	F_1: Not Confused	F_1: Confused	Kappa
Managing Emergencies	279 (18%)	0.963	0.771	0.741
Statistical Learning	3,030 (30%)	0.909	0.767	0.677
Economics 1	1,583 (23%)	0.933	0.741	0.675
Statistics in Medicine (2014)	1,218 (28%)	0.908	0.748	0.658
Statistics in Medicine (2013)	3,320 (21%)	0.916	0.671	0.589
Science Writing	5,181 (10%)	0.961	0.527	0.491
Women's Health	2,141 (15%)	0.933	0.506	0.445
How to Learn Math	9,878 (6%)	0.970	0.383	0.359

Table 3: Combined Confusion Classifier Performance, Individual Courses. Our classifier performed best on courses whose discourse was characterized by technical diction, like statistics or economics. In courses like *How to Learn Math* that facilitated open-ended and somewhat roaming discussions, our model found it more difficult to implicitly define confusion.

Humanities	Medicine	How to Learn Math	Managing Emergencies
constituent:urgency (6.59)	constituent:question (4.05)	constituent:question (6.64)	constituent:urgency (2.47)
constituent:question (3.47)	confused (2.98)	constituent:urgency (2.13)	constituent:question (2.34)
confused (3.20)	explain (2.71)	hoping (1.94)	? (1.73)
? (3.14)	role (2.41)	link (1.76)	metadata:#? (1.54)
couldn't (2.40)	understand (2.36)	available (1.63)	hope (1.40)
report (2.23)	stuck (2.27)	responses (1.62)	what (1.31)
manual (1.94)	struggling (2.25)	middle (1.62)	understand (1.29)
question (1.91)	constituent:urgency (2.25)	support (1.60)	dr (1.24)
haven't (1.84)	sentence (2.20)	discussing (1.60)	how (1.24)
wondering (1.83)	little (2.09)	instruction (1.60)	meant (1.23)

Table 4: Most Informative Features, Odds Ratios. Features prefixed with "constituent:" correspond to constituent predictions, while those prefixed with "metadata" correspond to post metadata features. All other features are unigram words.

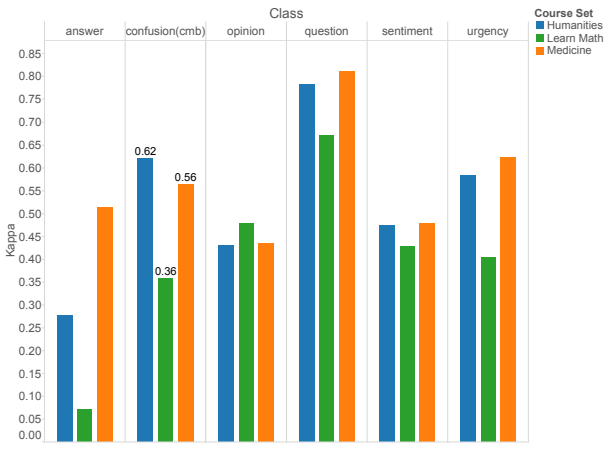

Figure 2: Constituent Classifier Performance. This figure visualizes the performance of the constituent classifiers, as well as the performance of the combined classifier (confusion(cmb), where "cmb" is short for combined).

Training Course	Test Course	Kappa
Stats. in Medicine (2013)	Stats. in Medicine (2014)	0.629
Stat. Learning	Stats. 216	0.590
Economics 1	Stats. in Medicine (2013)	0.267
Stats. in Medicine (2013)	Women's Health	0.175

Table 5: Nature of Confusion Across Domains. Training and testing on similar courses typically resulted in high performance, especially in the case of technical courses.

Figure 2 illustrates the performance of each constituent classifier when cross-validating on the humanities and medicine course sets, as well as on the education course. The constituent question classifier outperforms all the others by a large margin, likely because the structure of questions is fairly consistent. (Note that the constituent classifiers are not themselves fed by a lower level of classifiers; if we were attempting to predict, say, sentiment instead of confusion, we could try to improve over the performance shown here by creating a sentiment combination function that was informed by its own set of constituent classifiers.)

The combining fuction of our combined classifier consistently determined that the constituent classifiers for the question

Course Set	Humanities	Medicine	Learn Math
Combined Classifier	0.621	0.564	0.359
Minus Question	0.621	0.559	0.350
Minus Answer	0.620	0.555	0.345
Minus Opinion	0.619	0.553	0.310
Minus Sentiment	0.621	0.555	0.292
Minus Urgency	0.618	0.512	0.337
Minus Post Position	0.614	0.482	0.337

Table 6: Ablative Analysis, Kappas. Minus question is the combined classifier without the constituent question classifier; minus answer is the minus question classifier without the constituent answer classifier; and so on.

and urgency variables were particularly indicative of confusion (see Table 4). Table 6 shows the results of an ablative analysis in which one constituent classifier was removed from the combined classifier at a time, until we were left with a classifier with no constituent classifiers (call it a *flat* classifier). The flat classifier performed worse than the combined classifier in the two course sets and the education course, with the drop in performance most pronounced in the medicine course set. For both course sets, the urgency constituent seemed to be the most helpful of the five constituents – this makes intuitive sense, since we would expect that instructors would prioritize posts in which learners were struggling to understand the course material. However, the same was not true for *How to Learn Math*, which is consistent with the fact that no significant correlation between confusion and urgency was found (see section 3.2).

The post position metadata feature also contributed positively to the classifier's performance – removing it from the flat classifier for medicine dropped the Kappa by 0.03. The other metadata features, however, did not appear to consistently or appreciably affect classifier performance, and so we chose to omit them from our ablative analysis.

Table 4 shows that the number of question marks was also an informative feature, at least in the *Managing Emergencies* course.

6. PHASE II: RECOMMENDING CLIPS

6.1 The Recommendation Algorithm

In this section, we describe how YouEDU recommends instructional material for a forum post that has been labelled as *confused* by Phase I. Every course can be thought of as a collection several video lectures. Each video lecture on an average is about 12-14 minutes long. We focus on the problem of identifying a ranked list of snippets, S for each *confused* post. Each snippet s_i in S is a tuple *(video_id, seek_minute)* where *video_id* is an identifier for the recommended video and *seek_minute* is the time in the video to which the student must scroll and start playing the video. Since a concept is covered at multiple places in a video, we did not find it necessary to recommend an *end_minute*. Furthermore, students could end the video by themselves if they feel they adequately understood the particular concept.

Phase II of YouEDU is divided into an offline indexing phase and an online retrieval phase. We define a *bin* as a time-indexed section of a video. Each bin b_i can be thought of as the transcribed text content of the video at time interval i. We define *bin_score(w, b)* of a word w and bin b as the number of times word w appears in bin b. Our approach formulates video recommendation to learners as a classical Information Retrieval problem. In a classical IR task, the intent is to retrieve the top documents that match a user's query. In our case, the query corresponds to a confused post, and the document corresponds to a bin. We want to retrieve a ranked list of bins that express the content of the confused post. We describe the steps in Phase II of YouEDU in the following subsections.

6.1.1 Offline - Indexing Pipeline

Here are the steps involved in our indexing pipeline.

```
def build_index():
  bins = get_bins(course)
  for bin in bins:
  nouns = extract_nouns(bin)
  for noun in nouns:
   add (noun, bin) to the index
   term_weight[noun,bin]+=1
```

Binning: We hired translators to accurately transcribe the content of each video lecture into sub-title files. Each video has one corresponding sub-title file. In the indexing phase, *get_bins()* divides each video by time into several bins.

Pre-processing Bins: We use a part-of-speech tagger [**?**] to pre-process each bin. Nouns and noun-phrases tend to produce better key-words that typically express what the content is about [**?**]. Hence, we represent a bin as a triple (*video _id*, *start _min*, *noun _phrase _list)* where *noun _phrase _list* is a collection of only the nouns and noun-phrases in the bin.

Indexing Bins: We scan through each of the pre-processed bins, and build an index from each word to the corresponding bin that the word appears in. This index would enable us to retrieve the list of bins B_w, that corresponds to time epochs in the entire course the word w was discussed. We also maintain a data structure that keeps track of *bin_score(w,b)* for every word and bin. The constructed index and data structures are serialized to disk and is used by the retrieval phase which is done online.

6.1.2 Online - Retrieval and Ranking:

Here are the steps involved in our online retrieval phase.

```
def recommend(post):
 class = classify(post)
 if (class == confused):
  p = extract_nouns(post)
  candidate_bins = retrieve_candidates(post)
  for candidate in candidate_bins:
   b = vectorize(candidate)
   p = vectorize(post)
   sim_score = cosinesim(b,p)
  rank candidate_bins by sim_score
```

Pre-processing of forum post: We use a part-of-speech tagger [**?**] to pre-process each confused post. Similar to the

technique we used for bins, we represent each post as a list of its constituent nouns and noun-phrases.

Candidate bin retrieval: We scan through each of the words in the pre-processed post. We add bin b to the candidate set, if atleast one term in the pre-processed post was talked about in bin b. Since, we have the index constructed offline, we could use it to prune candidates from a large number of available videos (and hence, bins) in the corpus .

Candidate Ranking: We convert each post and bin into a V dimensional vector, where V is the size of the vocabulary computed over all words used in all lectures of the course. *vectorize(bin)* converts a bin into a V dimensional vector such that the value on the dimension corresponding to word w_i is *binscore(wi,bin)*. We define *simscore(P,B)* as the cosine similarity of the post and the bin.

$$simscore(P,B) = \frac{P.B}{\sqrt{\sum_{i=1}^{V} P_i^2}\sqrt{\sum_{i=1}^{V} B_i^2}} \quad (1)$$

For each candidate bin, C_i in the list of candidates C , we compute *simscore(C_i, post)*. We rank all bins in C by their *simscore* values and return the ranking.

6.2 Evaluation

We implemented our ranking system on a MOOC on 'Statistics in Medicine' offered here at Stanford University that had 24943 learners. Our classifier to detect confusion in forum posts was trained on one run of the course and tested on the other run. We chose a random sample of queries from our MOOCPosts dataset for that course. We ran each of those posts through Phase I of YouEDU. We chose 20 random posts from the posts that were labeled as confused. For each of those *confused* posts we obtained a list of six ranked video recommendations. We then requested three domain experts (in statistics) at Stanford to rank and assess the quality of the recommendations with respect to each confused post.

The rating scale is described below

Type1: Relevant (the recommended snippet is exactly relevant to the question in the forum post and addresses the content of the post precisely)

Type2: Somewhat relevant (the recommended snippet is relevant to the post and is useful to address the learner's question in the post)

Type3: Not Relevant (the recommended snippet does not address the specific question in the forum-post).

We describe the two metrics that we use for evaluating how relevant our recommendations are and how much they agree with the gold-set.

Normalized Discounted Cummulative Gain (NDCG): NDCG measures ranking quality as the sum of the relevance scores (gains) of each recommendation. However, the gain is discounted proportional to how below the document is in the ranking. The underlying intuition is that the gain due to an extremely relevant document (say, relevance score of 2) appearing as the last result must be penalized more than if it appeared at the first result. Hence, the DCG metric applies a logarithmic discounting function that progressively reduces a document's relevance score (gain) as it's position in the ranked list increases [**?**]. The base b of the logarithm measures how sharp the applied discount is.

Rater	NDCG	k-precision k=1	k=2	k=3
Rater1	0.66	0.66	0.61	0.62
Rater2	0.90	1.0	0.97	0.97
Rater3	0.82	0.55	0.52	0.52
Avg	0.79	0.74	0.70	0.70

Table 7: NDCG and k-Precision for recommendations

If rel_i is the gain associated with document at position i, The DCG at a position i is defined recursively as

$$DCG(i) = \begin{cases} rel_{\mathrm{i}} & i \leq b \\ DCG(i-1) + \frac{rel_i}{\log_b i} & otherwise \end{cases} \quad (2)$$

Since we want a smoother discounting function, we set b to 2. We used a graded relevance scale of 0, 1 and 2 (2 corresponding to most relevant and 0 corresponding to irrelevant) and computed the DCG for the ranked recommendations we obtained for each confused post. The Ideal value of DCG (IDCG) is defined as the DCG based on the ideal ranking calculated by the raters. To obtain the IDCG, we sort the rankings given by the raters in decreasing order of relevance scores and compute the DCG of the sorted ranking. This corresponds to the maximum theoretically possible DCG in any ranking of recommendations for that post. We normalize the DCG for our ranking by the IDCG to get the Normalized Discounted Cummulative Gain (NDCG).

$$NDCG(i) = \frac{DCG(i)}{IDCG(i)} \quad (3)$$

Precision at top k: We define Precision of a ranking R with n recommendations as the fraction of the recommendations that are relevant.

$$Precision(R) = \frac{\#\text{ of Relevant Recommendations}}{n} \quad (4)$$

The Precision at k of a ranking R is defined as the value of the precision considering only the top k recommendations in R.

Our results across the raters are summarized in Table 7. Our average precision at k=1, is 0.74. This intuitively means that on about 74% of the cases, the first video that we suggest to a learner (as a recommendation for his confused post), is a relevant video. We think this is a remarkable precision for a completely automated intervention system. Our precision values for k=2 and k=3 are also encouraging. Our NDCG numbers are high indicating that we perform relatively well to the theoretically maximum possible DCG (ie, DCG of a ranking computed by a human domain expert). We don't consider recall into our evaluation as there are only very few available videos for most confused post. So, the problem of missing out on videos does not arise.

7. FUTURE WORK

The work we presented here is a first step. Many opportunities for improvement remain. We are actively investigating whether we can strengthen our video snippet ranking further by considering which video portions learners re-visited several times. This analysis will catalog the number of views that occurred for each second of each instructional video in a course. The hope is that re-visit frequency could further boost our recommendation ranking algorithm.

Another thrust of future work will use the question and answer classifiers to actively connect learners to each other. The challenge to meet in this work is to identify learner expertise by their answer posts, and to encourage their participation in answering questions related to their expertise. As in this work, auxiliary data, such as successful homework completion, will support this line of investigation.

A third ongoing project in our group is the development of user interfaces for both instructors and learners. Using our classifiers, we have been experimenting with interactive visualizations of our classifiers' results. The hope is, for example, to have instructors see major forum borne evidence of confusion in a single view, and to act in response through that same interface.

Video recommendations are not the only source of help for confused learners. Many online courses are repeated during multiple quarters. It should therefore be possible for our system to search forum posts of past course runs for answers to questions in current posts. Also, not all confusion is resolvable through the videos. For example, difficulties in operating the video player is unlikely to have been covered in the course videos. Distinguishing such posts is an additional challenge.

YouEDU's two phases need not be packaged together; in an online setting, they could operate as independent, complemetary services. The output of Phase I could be presented directly to instructors, many of whom express interest in understanding activity in discussion forums [**?**]. As for Phase II, the recommendation system might be made accessible through a search interface: instead of posting, learners would type natural language queries in which they voiced their confusion, and the apparatus presented here could serve relevant resources.

As novel online teaching methods are developed, the same underlying challenges will need to be met: keeping learners engaged, allowing them to feel like members of a community, and maximizing instructor effectiveness in the difficult environment of large public classes. More questions will therefore arise as online teaching methodologies change.

8. CONCLUSION

We presented our two phase workflow that in its first phase identifies confusion-expressing forum posts in very large online classes. In a second phase, the workflow uses those post texts to identify related excerpts from instructional course videos. These minute resolution snippets are then offered as recommendations to the confused learner.

Our approach utilizes new data sets of human tagged forum posts, data from learner interactions with online learning platforms, and video closed caption files that are produced in concert with the videos for vision impaired learners.

We evaluated the classifiers that identify posts from confused learners, the effectiveness of subsequent video snippet retrieval, and the ranking process that orders those snippets into ranked lists.

We also pointed to significant amounts of research work still to be accomplished. The challenge of teaching online to very large numbers of learners from diverse backgrounds remains formidable. But the potential benefits to large underserved populations should encourage the required investigative effort.

9. ACKNOWLEDGMENTS

This section is optional[**?**]; it is a location for you to acknowledge grants, funding, editing assistance and what have you. In the present case, for example, the authors would like to thank Gerald Murray of ACM for his help in codifying this *Author's Guide* and the **.cls** and **.tex** files that it describes.