Weakly Supervised Learning of Dialogue Structure in MOOC Forum Threads

Robert Fisher
Carnegie Mellon University
Pittsburgh, PA 15213
Email: rwfisher@cs.cmu.edu

Reid Simmons
Carnegie Mellon University
Pittsburgh, PA 15213
Email: reids@cs.cmu.edu

Caroline Malin-Mayor
Brown University
Providence, Rhode Island 02912
Email: caroline_malin-mayor@brown.edu

***Abstract*—In this paper we present a new method for understanding discussions between students in MOOC forums. In particular, we introduce a machine learning method for discovering instances in which a response relation exists between a pair of posts in a forum thread, for example when one student provides the answer to a question or comments on something another student previously said.**

Research has shown that understanding conversational structure between students is paramount to evaluating the productivity of the collaboration and estimating outcomes. However, previous methods often rely on human supplied dialogue act labels or discourse parsing algorithms requiring large labeled datasets. Our method, which utilizes a fast, exact optimization process known as *spectral optimization*, does not require manually annotated training data and is highly scalable and generalizable. Empirical results are given using real world datasets consisting of conversations between students participating in Coursera courses, and we see predictive accuracy above 90%–nearing the human inter-annotator agreement rate for these datasets.

I. Introduction

Massive Online Open Courseware (MOOC) offers tremendous potential for expanding access to quality education. However, as the number of students vastly outweigh the number of administrators and educators, it is vital that productive collaboration between students is fostered. Given the sheer volume of forum activity in a large MOOC, there is an opportunity to develop software techniques that can automatically help to analyze natural language discussions between students. This analysis can be used to clean and curate course forums, as well as deploy automated interventions to aid struggling students. We may, for instance, identify which students are most actively engaged in productive conversations with their fellow students, and these peer mentors could be leveraged when another student is found to be struggling.

Shallow, sentence level language analysis has proven useful in a variety of MOOC applications, however many of these methods ignore the higher level inter-sentential structure of online discussions. There is an opportunity to better understand conversations between students by identifying which posts within a thread are related to one another, which induces a directed relationship graph for the thread. Conversational structure, such as dialogue act sequences, can be very useful when assessing the quality of unstructured discourse between students [1]. Unfortunately, most of this work has relied on manually annotated conversational structure. In this paper, we present a method to automatically discover forum thread structure in a weakly supervised manner that does not require costly, manual annotation of data. Instead, we rely on training signals inferred from a thread's metadata, or supplied directly by the students during normal forum usage.

Discourse parsing offers a useful framework for analyzing the high-level structure of a conversation or a block of text. In most discourse parsing work, semantic relationships are identified between *elementary discourse units* such as sentences or sentence fragments. This framework generally requires very large manually annotated datasets, such as the Penn Discourse Treebank [2]. Compiling this sort of annotated corpus is difficult for MOOCs, where supervised classifiers may not generalize across different course topics. Instead, we consider a simpler problem in which we seek to identify pairs of forum posts in which the second post is a direct response to the first. This allows us to leverage structured forum responses built into popular online platforms such as Coursera. Coupling these structured responses with a semi-supervised training framework allows us to train a discourse parser that does not require manually supplied training data.

As with any discourse parsing task, individual subparts of the discourse structure can depend strongly on the global characteristics of the tree, and these dependencies have been ignored by many previous approaches. Modeling these global dependencies explicitly would require computing the joint probability distribution of the whole structure, which is generally intractable. To overcome this obstacle, latent variable models have become a popular tool for many structure learning tasks. In this framework, we assume the existence of Markovian hidden states attached to the nodes or edges of the learned structure, which allow us to model global dependencies without the need for a complex and sparse joint distribution. Even with these simplifying assumptions, learning the parameters of the latent variable model is a difficult task that is often solved with heuristic or suboptimal optimization.

There has recently been growing interest in a breed of algorithms based on spectral decomposition. Spectral algorithms utilize matrix factorization algorithms such as Singular Value Decomposition (SVD) and rank factorization to discover *low-rank decompositions* of matrices or tensors of empirical moments. In many models, these decompositions allow us to identify one or both of two things: first, the subspace spanned by a group of parameter vectors, and second, the actual parameter vectors themselves. For tasks where they can be applied, spectral methods provide statistically consistent results that avoid local maxima. And, spectral algorithms tend

to be much faster–sometimes orders of magnitude faster—than competing approaches. These methods can be viewed as inferring something about the latent structure of a domain–for example, in a hidden Markov model, the number of latent states and the sparsity pattern of the transition matrix are forms of latent structure, and spectral methods can recover both in the limit.

In this paper we present a spectral, hidden variable parsing algorithm for learning the relational structure of conversations between students participating in a MOOC, which is based on a framework that has previously been used for discourse relation classification [3]. We describe the specifics of the Coursera data, consisting of forum activity taken from two courses, and we provide empirical results on a human annotated testing set compiled from the data. The model is able to achieve predictive accuracy of .921 and .903 on the two course datasets, which is nearing the estimated human inter-annotator agreement of .939.

II. Related Work

Education research has demonstrated that characteristics of dialogue and discourse between students can be used to evaluate the effectiveness of the discussion and predict student outcomes [4], [5], [1]. However, these approaches often either rely on human supplied dialogue tags, or fully supervised discourse parsers that require large, labeled datasets. As such, these systems show great promise on the domains in which they are deployed, but they are often not generalizable.

In the field of computational linguistics, existing algorithms for learning discourse structure rely on slow or inexact optimization procedures. Recently, efforts have turned towards Conditional Random Fields (CRFs), showing significant improvements compared to simpler methods [6]–but traditional CRFs continue to rely on slow optimization that is susceptible to local minima. Thus, there is potential for spectral methods to advance the study of discourse parsing through a more complex but still time efficient model.

Spectral methods have been successfully employed in many applications, including knowledge tracing for education [7]. There has also been work training a latent variable model for dependency parsing using EM [10], but this approach was slightly less accurate and orders of magnitude slower than the spectral equivalent. Additionally, there has been quite a bit of work concerning the use of spectral methods for supervised models [11], [12] and latent variable models with known structure [13], [14]. Semi-supervised, spectral discourse parsing was first introduced in [3], but this work did not address structure learning, relying instead on discourse sequences taken from the Penn Discourse Treebank. This work demonstrated that a semi-supervised, spectral method outperforms fully-supervised discourse parsing approaches such as SVMs and CRFs.

III. Dataset

The dataset for our experiments consists of forum posts from two Coursera courses. Some statistics describing these datasets are shown in table I. The first course, 'Learn to Program: The Fundamentals', focused on teaching the Python programming language to students with little computer science experience. The second dataset, based on a psychology course, was used in the shared task for the Modeling Large Scale Social Interaction in Massively Open Online Courses Workshop [15]. Posts are organized into threads, and each post is accompanied by a set of metadata variables, including the author of the post, the title of the thread, the timestamp, the number of votes given to the post by other students, as well as a tag indicating if the thread the post is contained in has been identified as resolved by the thread's creator. Additionally, each post is designated either as an unstructured member of the forum thread, or as a *comment* which is a structured method for forum participants to respond to another specific post.

	Python	**Psychology**
Total Threads	3,234	1,488
Total Posts	24,963	5,244
Mean Posts/Thread	7.7	3.5
Total Comments	8,421	2,747
Labeled Threads	54	34
Labeled Pairs	1,120	916

TABLE I. MOOC Datasets

A subset of the dataset was manually annotated for use in testing the model's predictive accuracy, and the sizes of these testing sets are also shown in table I. It should be noted again that this labeled data is only used for evaluation, and no manually annotated data is required for training. Each pair of posts in this set was evaluated to determine if the second post in the pair was a direct response to the first, and a label was given for every ordered pair of posts taken from the threads selected for annotation. Five threads from the psychology dataset were labeled by two different annotators, and within those threads we observed an inter-annotator agreement rate of .939. Of all possible pairs of posts in the testing sets, 24.7% of them were observed to share a response relation in the Python course, and 19.1% were related in the psychology course. Structured comments were seen to be related to their parent post roughly 60% of the time, while the second post in a thread was observed to be a direct response to the first post 94% of the time. For the purposes of evaluation, only unstructured pairs of posts were used for testing. Structured comment posts and the first two posts in every thread were excluded.

IV. Method

In this work, we use Markovian latent states to compactly capture global information about a parse sequence, with one latent variable for each relation in the discourse parsing sequence. Most other discourse parsing frameworks label each relation independently, but our model allows information about the global structure of the discourse parse to be used when identifying an individual relation.

Our model considers each post to be one elementary discourse unit. Each thread of length t produces $t-1$ sequences of posts. Specifically, we first consider all pairs of posts in the thread of distance 1 in chronological order, *i.e.* all adjacent posts. We then consider all pairs of post of distance 2, etc. This allows us to identify all responses in a sequence, not just relations between adjacent discourse units as seen in datasets such at the Penn Discourse Treebank. This is especially important for forum posts which are asynchronous, which may often result in relations between non-adjacent posts in a thread.

According to our assumption of Markovian structure, each potential relation x_t between elementary discourse units *edu*$_t$ and *edu*$_{t+1}$ is accompanied by a corresponding latent variable h_t, and conditioning on h_t makes x_t independent from $x_{1...t-1}$ and $x_{t+1...n}$. This structure allows us to use an HMM-like latent variable model based on the framework presented in [11], [16].

Since the typical parameterization of an HMM (consisting of an initial state distribution $\vec{\pi}$, a state transition matrix T, and an observation matrix O) is difficult to learn directly, we use the *observable* formulation of the model. Define the *observable operator* matrix A_x as

$$A_x = T\text{diag}(O_{x,1}...O_{x,m})$$

Then the following equality holds:

$$Pr[x_1...x_t] = \vec{1}_m^\top A_{x_t}...A_{x_1}\vec{\pi}$$

Using this formulation of the model, we need to learn the matrices A_x and T, as well as the vector $\vec{\pi}$. To compute these parameters, we use the unigram, bigram, and trigram probability matrices. We denote the unigram matrix as P_1, the bigram matrix as $P_{2,1}$, and the trigram matrix as $P_{3,x,1}$, with one trigram matrix for each value of x. Define these matrices as follows:

$$[P_1]_i = Pr[x_1 = i]$$

$$[P_{2,1}]_{ij} = Pr[x_2 = i, x_1 = j]$$

$$[P_{3,x,1}]_{ij} = Pr[x_3 = i, x_2 = x, x_1 = j]\ \forall x$$

Spectral latent variable models utilize subspace identification to learn the model dynamics in a reduced dimensionality space. If we assume that the dimensionality of this subspace is no less than the rank of the parameter matrices O and T, then the model learned in the subspace is equivalent to the model from the original feature space. There are different methods of projecting the data into this subspace, but one convenient approach is to take the left Singular Values of the bigram matrix $P_{2,1}$. We denote this subspace transformation matrix as $U \in \mathbb{R}^{n \times m}$. If we denote the subspace model parameters as π_U, T_U, and A_U, these parameters can be easily learned using factorization over the fully observable matrices defined above. Proofs of all equalities given in this section are available in [11].

$$\hat{\pi}_U = U^T P_1$$

$$\hat{T}_U = (P_{2,1}^T U) + P_1$$

$$\hat{A}_U = U^T P_{3,x,1}(U^T P_{2,1})^+\ \forall x$$

For the feature space, we use the rich linguistic discourse parsing features defined in [17], which include syntactic and linguistic features taken from dependency parsing, POS tagging, and semantic similarity measures. The various metadata for each post is also included as features, as is the time between publication for each pair of posts, and the distance between their positions in the forum thread. We also include a feature indicating if the name of the author from the first post appears in the body of text in the second post.

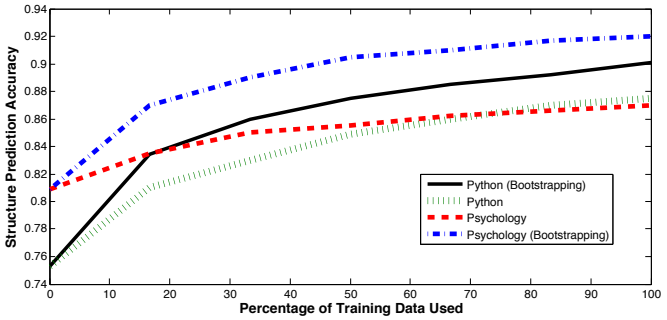

Fig. 1. Empirical Prediction Results Using Coursera Data

In addition, a Latent Dirichlet Allocation (LDA) model is trained using the English language Wikipedia, and this model is used to compute a vector space representation of the semantics of each post. From the LDA model, the top 400 concepts for each post are included in the feature space, as is the cosine similarity of the vector space representations of the two posts. URLs, HTML tags, and Python code are removed from posts before the vector space representation is computed. LDA concepts related to math, science, and technology were more common in the Python dataset, while the psychology dataset included more concepts related to the social sciences and popular culture.

We use one variable in the observation space to denote the response relationship for a pair of posts, indicating whether there should be an edge between these posts in the discourse tree. If the second post in the pair is a comment directed toward the first post, we consider this a responsive pair and set the response variable to 1. Initially, all structured comments are labeled as related to their parent post, as are the first two posts in every thread. Together, these two sets of datapoints are used to train an initial model. For all other pairs in the dataset, the response variable is set according to the model's estimated probably that the pair constitutes a response relation given the other observed variables, and the latent state distribution. Initially this probability is set to the naive prior, but the estimates for the unlabeled data change every time the parameters are recomputed. Unlabeled pairs with a predicted relation probability surpassing a threshold ϕ have their labels set to 1. Because of the speed of the spectral method, we are able to construct estimates for the response variables and recompute the model parameters several times.

After five iterations of bootstrapping, the classifier is then used to predict the relationships of all structured comments with their parents. Those comments with a predicted confidence below a threshold, ψ, become negative training samples, and the model is trained one more time. This step accounts for the fact that many students use the comment feature in unintended ways, which results in many false positive data points being added to the training set. A small parameter validation set was held out separately from the training set and used to select parameters ψ and ϕ.

V. Results

Figure 1 shows empirical results using the spectral HMM to predict response relations in the Coursera dataset as a function of the percentage of training data used. The training curves denoted with (Bootstrapping) include the stages in which the unlabeled data is incorporated into the training set and the labels of the structured comments are corrected. The other training curves include neither unlabeled data nor label correction.

We see a maximum classification accuracy of 0.921, and a maximum precision of 0.782. This predictive accuracy is nearing the estimated, human inter-annotator agreement of .939, which can be thought of as an upper-limit on predictive performance at this task. Without bootstrapping, when the model has only used structured comments and the first two posts of every thread for training, the maximum precision drops to 0.689. This indicates that leveraging unlabeled data and relabeling incorrect comments results in a 10 percentage point increase in prediction precision. For comparison, a naive classifier that predicts that no pairs of posts share a relation would produce an accuracy of 0.638 with 0 precision. We also note that the psychology dataset yields a higher overall accuracy but a lower precision. The decrease in precision is likely due to the smaller size of the training set, while the increase in accuracy may simply be a consequence of the distribution being more skewed towards negative examples, leading to a higher baseline prior.

In these experiments, the HMM parameters were computed and used to produce estimates for the relation probability of all pairs in the training data that did not have a comment relationship. These estimates were then used to recompute the HMM parameters, and this process was repeated five times. The labels of the structured comments were then predicted, and all comments with a relation probability below ψ became negative training results, while unlabeled pairs with a confidence above ϕ became positive training examples.

If the labels of the structured comments are not corrected after bootstrapping, the precision in both datasets drops by approximately 4 percentage points. Using the value of ψ chosen with the validation set, we observe that roughly 30% of the structured comments have their labeled switched to negative. Figure 1 does not include results with comment prediction, but separate tests suggest the predictive recall for comments is nearly 88%. This indicates that the corrective step of the algorithm could be reducing the percent of false positive training samples from 40% to as low as 15%.

VI. Conclusion

In this paper we have presented a spectral, weakly supervised method for learning the inter-post structure of online forum threads. Previous research has shown many applications in which discourse information is useful such as dialogue evaluation, thread resolution prediction, and thread curation. There are also other applications that have not been widely studied. For instance, the response relations between posts tell us explicitly which students are speaking to one another in the forums. This could help us identify potential peer-mentors that commonly answer questions for their fellow students, and these mentors could be leveraged by an automated intervention to minimize the workload for paid teaching assistants.

An opportunity for future work is to create new weakly supervised methods to learn multiple semantic relations between posts. While fully unsupervised methods might not reliably learn the most useful types of relations, semi-supervised methods or techniques that leverage existing structured data show promise. Regardless of the methods used, the sheer scale of MOOCs presents a fantastic opportunity for data-mining in a variety of tasks, but many existing language analysis methods are often bottlenecked by the need for annotated data. By better utilizing vast stores of unlabeled data, we can build tools to benefit educators and students alike.

References

[1] D. Adamson, A. Bharadwaj, A. Singh, C. Ashe, D. Yaron, and C. P. Rosé, "Predicting student learning from conversational cues," in *Intelligent Tutoring Systems*. Springer, 2014, pp. 220–229.

[2] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber, "The penn discourse treebank 2.0." in *LREC*. Citeseer, 2008.

[3] R. Fisher and R. Simmons, "Spectral semi-supervised discourse relation classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, vol. 2, p. 89.

[4] B. Samei, A. Olney, S. Kelly, M. Nystrand, S. D'Mello, N. Blanchard, X. Sun, M. Glaus, and A. Graesser, "Domain independent assessment of dialogic properties of classroom discourse," in *Educational Data Mining 2014*, 2014.

[5] E. Mayfield, D. Adamson, and C. P. Rosé, "Hierarchical conversation structure prediction in multi-party chat," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2012, pp. 60–69.

[6] V. W. Feng and G. Hirst, "A linear-time bottom-up discourse parser with constraints and post-editing," in *Proceedings of The 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, USA, June*, 2014.

[7] M. H. Falakmasir, Z. A. Pardos, G. J. Gordon, and P. Brusilovsky, "A spectral learning approach to knowledge tracing," 2010.

[8] G. A. Musillo and P. Merlo, "Unlexicalised hidden variable models of split dependency grammars," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, 2008, pp. 213–216.

[9] D. Hsu, S. M. Kakade, and T. Zhang, "A spectral algorithm for learning hidden markov models," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1460–1480, 2012.

[10] B. Boots and G. J. Gordon, "Predictive state temporal difference learning," *arXiv preprint arXiv:1011.0041*, 2010.

[11] L. Song, E. P. Xing, and A. P. Parikh, "A spectral algorithm for latent tree graphical models," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 1065–1072.

[12] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y.-K. Liu, "A spectral algorithm for latent dirichlet allocation," *arXiv preprint arXiv:1204.6703*, 2012.

[13] D. Yang, M. Wen, and C. Rose, "Towards identifying the resolvability of threads in moocs," *EMNLP 2014*, p. 21, 2014.

[14] B. Boots, S. Siddiqi, and G. Gordon, "Closing the learning-planning loop with predictive state representations," *Intl. J. Robotics Research (IJRR)*, vol. 30, no. 7, 2011.

[15] V. W. Feng and G. Hirst, "Text-level discourse parsing with rich linguistic features," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 60–68.