Prediction of Undergraduate Students' Course Sequences and their Naturalness

Stanford CS224N Custom Project

Shyamoli Sanghi
Stanford University
shyamoli@stanford.edu

Victoria (Docherty) Delaney
Stanford University
vdoch@stanford.edu

Qi Han
Stanford University
qihan96@stanford.edu

Abstract

Undergraduate students' success in coursework is associated with persistence along a degree trajectory. The ability to predict future courses given a past sequences of courses is, therefore, of great value to educators and universities, as it can help degree-seeking students select which courses to take and when to take them. To add to qualitative findings from the major forecasting field, the following study uses LSTM and RoBERTa models to predict students' future courses given their freshman course histories, as well as make predictions on the degree of naturalness of a freshman course trajectory. To accomplish these tasks, we drew upon 26,892 undergraduate students' enrollment decisions data from Carta, a Stanford web-based course exploration tool. Students freshman-year courses and grades were represented as one-hot encodings and modeled at the character level. Results suggest that the pre-trained RoBERTa model was substantially more accurate (89.01%) in future course predictions, that bidirectional representation of course names is important for course prediction tasks, and that students naturally enrolled in courses in a larger number of distinct subjects than in a lower number of distinct subjects during their freshman years. These findings, while consistent with students' course exploration trends at liberal arts schools, merit future exploration to determine the impact of students' grades on their decisions to explore and commit to a degree trajectory.

1 Key Information to include

- Mentor: Kathy Yu

2 Introduction

Course selection is a high-stakes endeavor for college students. Students' experiences in early, freshman-year courses are likely to determine a number of future outcomes, including academic success, satisfaction, and persistence along a degree trajectory [1]. However, many universities (particularly, liberal arts schools) encourage exploration and breadth in student coursework as core to their academic program. Given students' need to balance exploration with persistence to complete degree programs, this study examines the composition of freshman-year course histories at a medium-sized liberal arts university and leverages them to predict future (sophomore-year) courses, used as a proxy for academic trajectory persistence.

Though degree trajectory forecasting and course predictions have been explored from sociology perspectives ([2], [3]) and through various quantitative, language-modeling techniques ([4]), results from these studies are often presented at a large grain size. Our approach builds on course prediction as a language modeling task and differs from prior studies by (1) measuring the "naturalness" of students' course histories, which measures the likelihood that a student would take on a particular

course history, (2) by taking into account grade-level performances in students' freshman year courses, and (3) by making course predictions at the character level rather than by words or subwords.

To accomplish this task, we developed character-level course and grade embeddings and compared course prediction performance between a stacked LSTM model and a fine-tuned RoBERTa (citation to the RoBERTa paper) encoder model. Training at the character-level afforded the opportunity to examine model accuracy in course prefixes (e.g., "CS") and course levels (e.g., "CS124" vs. "CS224n"). Through this approach we address the following research questions:

- What improvements in course-prediction accuracy can be made by modeling and predicting students' future courses (1) at the character-level and (2) using grade-level embeddings?
- What can we learn about the naturalness (likelihood) of freshman year course sequences and the natural number of subjects taken?

3 Related Work

3.1 Pre-training Models' Inspirations: BERT and RoBERTa

Large encoder models, such as BERT, and RoBERTa, serve as the foundation for our study's overarching task. Both are pre-trained on masked language modeling and next-sentence prediction, which corresponds to the course sequences and course characters this paper aims to analyze. BERT [5], Bidirectional Encoder Representations from Transformers, is the original conceptual model for this project because it achieves state-of-the-art performance on both sentence-level and token-level tasks. Its use of a Masked Language Model(MLM) pre-training objective fuses contexts on the left and right sides of masked tokens in a deeper representation level than other unidirectional models. Its bidirectionality also helps integrate adjacent courses' information, which facilitates the pre-training process on a character level.

Our experiment imports RoBERTa [6] as it improves the performance of BERT on GLUE (General Language Understanding Evaluation) tasks by modifying some design choices. Since RoBERTa was pre-trained with a 160-GB, uncompressed English data set, the transformer already contains some information features of the characters. This coverage could help facilitate connections among characters, such as understanding the string "CS" as it semantically relates to "Computer Science" instead of a simple concatenation of letter "C" and "S."

3.2 Forecasting Students' Academic Trajectories with Previous Courses

Our work shares a data set with and builds from the findings of Lang and colleagues [7], who used word vector embeddings to represent courses and a shallow-learning algorithm to forecast student majors. Lang and colleagues reported that students' first 25 courses predicted their eventual major thirty times better than random guessing. We aim to create deeper representations of course names using RoBERTa and hope to further improve the accuracy of course predictions using character-level specificity.

Using neural architectures to predict course sequences, make recommendations, and forecast majors are not uncommon. Shao, Guo, and Pardos [4] built upon BERT's encoder base and introduced PLAN-BERT, which addresses a need for multiple consecutive course recommendations (e.g., a term's worth of courses). A number of studies have also incorporated students' grades into prediction models, although these studies have predicted grades as outcomes rather than utilized grades as input embeddings [8] [9]. Regardless, this growing body of literature suggests that forecasting academic trajectories is a relevant and timely problem space given the predictive power of encoder models.

4 Approach

Task Summary: We aim to do the following:
(1) to predict future course, given sequence of past freshman year courses (+sequence of past grades)
(2) to predict the negative log-likelihood (cross-entropy loss) of a given course sequence, which measures the degree of "naturalness" of a sequence, as well as the number of distinct course subjects taken in freshman year.

Data-cleaning and pre-preprocessing: We were given access to the data by Stanford's Carta Decision Pathways Lab in a 1.5 million-row, 37-column CSV file, organized by student-enrollment decision pairings. 33,257 students over 20 years (2000-2020) are represented in the data. The original file contained extraneous columns that were not directly related to our research question. As such, we eliminated all but 6 columns and converted them into a 1,500,000 by 6 data frame using the Pandas library. The retained columns contained (1) a unique student identifier, (2) the course enrollment term and year, (3) the course subject and (4) catalog number, (5) the student's final grade in the course, and (6) the student's degree plan. We concatenated the course subject and catalog number to support our models' character-level embeddings where both letters and numbers are significant; for instance, "PEDS" and "216" became "PEDS216".

For the data to be pliable with our baseline model, we organized it into a master list of 33,257 dictionaries, where each dictionary represented one student's enrollment history. Each dictionary contained four keys: 'course-year', paired with a list of all years when courses were taken; 'course-list', paired with a list of all courses taken by the student; 'grade-list', paired with a list of all grades received by the student; and 'degree-plan', paired with the degree declared by the student. All lists were constructed such that order was preserved. For example, if a student's dictionary contained the following: Course-Year: [2019, 2018...], Course-List: [SOC102, PEDS216...], Grade-List: [A, B+, ...], this would indicate that the student took SOC102 in 2019 and received an "A", took PEDS216 in 2018 and received a B+, and so forth. Students represented in the data frame with empty dictionaries were removed. Finally, we extracted students' first sophomore-year courses as labels for the course prediction task.

Feature Representation: We created sequences of courses where each course represented a Course subject + Course Catalog Number (for instance, "CS" + "106A" -> "CS106A"). We then created one-hot encodings for each character in the sequence of course catalog numbers. In order to convert the characters to numbers, we created a char2int dictionary that mapped each character to its corresponding integer value. We modeled features at the character level because the one-hot embedding dimensions would be smaller and less sparse than course-level one-hot embeddings.

To make the feature representation of each student uniform in length, we found the maximum number of characters in all students' course sequences and padded the other course sequences to equal the length of the maximum course sequence (252 characters). Our character space contained 39 possible characters: 26 English letters, 10 numeric digits, 1 comma, an '' symbol, and 1 '%' padding character.

In order to include information about students' grades in freshman courses, we created one-hot embeddings representing grades that students earned in each course in their course list. There were 49 possible grade options: ('CR', 'S', 'C+', 'B+', 'C-', 'B', 'B-', 'A', 'A-', 'D', 'C', 'A+', '+', 'NC', '*', 'W', 'RP', 'HP', 'P-', 'NP', 'L', 'D-', 'D+', 'GNR', 'H', 'P', 'MP', 'N', 'I', '3.3', '3.4', '2.9', 'LP', 'U', 'P+', '3.6', 'N-', '4.1', 'KM', 'K', '-', '3.7', '4.3', '4.0', '3.5', '3.2', '3.9', '3.1','%'). Thus, we created each grade representation to be a 49-component, one-hot encoding.

Baseline Model: Stacked LSTM with Course Sequence Embeddings.

We created a baseline model: a character-level sequence model that comprised of 2 LSTM layers and one Linear layer. The first LSTM layer outputted a hidden state with dimensions that represented (sequence length)*(first hidden size) where we set return sequences = True. The second LSTM's hidden state had dimensions equal to the second hidden size and again, we set return sequences = True. This was smaller than the first hidden state's size. We used ReLU activation functions between these layers. Further, we employed the Adam optimizer and used the Cross-Entropy Loss function as our training objective. Importantly, our approach and implementation were completely original. The feature representations and baseline models were generated entirely from our own efforts.

Stacked LSTM with Course Sequence and Grade Embeddings. We created a grade embedding matrix for each student by finding the maximum number of courses taken (and hence, grades received) by any student in the training data. Then, we padded all other embeddings such that the sequence length was uniform. Thus, the dimensions of this matrix were (number of students, max number of courses, grade embedding dimension). Since we had to concatenate this matrix with the course sequence input matrix and wanted the first and second dimensions to match, we padded the grade

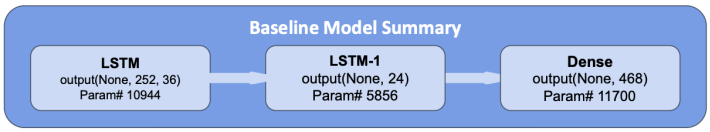

Figure 1: Baseline Model Summary

embeddings further such that the second dimension represented the maximum number of characters in any course sequence. Then we concatenated the grade embedding matrix to the course sequence embedding matrix in the last dimension. Finally, we used Truncated SVD to extract the most important 39 components of the last dimension, so as to reduce the last dimension size to the embedding size of the course sequence matrix (since this was the shape required by the model).

Encoder Model: In accordance with the second aim of this project, we wanted to use a technique to measure the naturalness of a freshman course sequence. Thus, we used a pre-trained language model and fine-tuned it on our data, as our data consists of courses that are different than English language words that language models are pre-trained on. Since RoBERTa was trained on Masked Language Modelling and Next Sentence Prediction tasks, we used it to determine the negative log-likelihood through the cross-entropy loss of a course sequence. This, in turn, allowed us to determine the naturalness score of each course sequence.

5 Experiments

5.1 Data

The data are described in the section above. Before running the experiments, we randomly sampled 90% of the data as our training data and left the remaining 10% as the test data. Data for the grade-level LSTM model were split into batches of 128 in order to be processed by Google Colab, as the high dimensionality of the embeddings matrix made modeling with the entire data set computationally intractable.

5.2 Evaluation Method

For the two LSTM experiments, we compared the number of characters that matched ground truth outputs and hence computed character-level accuracies of model outputs. Thus we defined the following evaluation metrics:

Fraction of Matching Characters: We compared the fraction of common characters in the model output course sequences and their corresponding ground truth labels.

Character-Level BLEU Score: We computed the BLEU scores of each of the models' output course sequences, where we took a course sequence to represent a sentence of characters. This took into account the number of matched character-level n-grams in model outputs and data labels, just as a character-level BLEU score would.

We also used course-level accuracy and negative log-likelihood(cross-entropy loss) as measures of accuracy of course prediction and unnaturalness (unlikeliness) of a course sequence, respectively.

5.3 Experimental Details

For the Baseline LSTM Model and LSTM with grade embeddings, we used ReLU activation functions between both hidden layers, the Adam optimizer, as well Cross-Entropy Loss as our objective function. While the number of training examples (students) was 33,196, the number of validation examples was 3,329. Batches of the size of 128 for the training process in the grade-level LSTM due to the computational expenses imposed by high-dimension embeddings.

The plots below represent the training loss over time for the Baseline LSTM model (see Figure 2) and for the LSTM model with grade embeddings (see Figure 3) respectively:

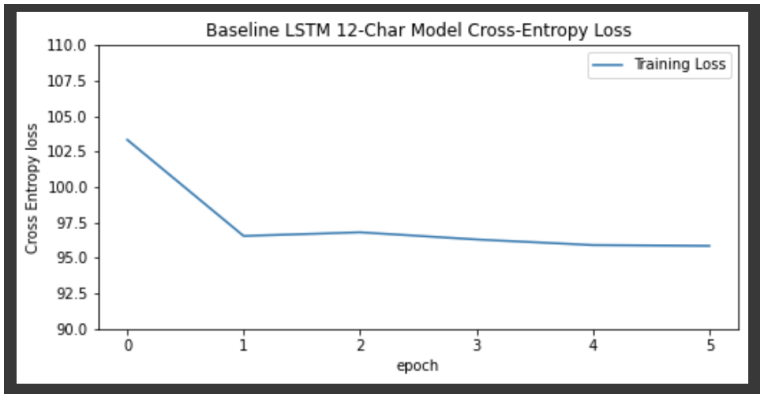

Figure 2: Baseline LSTM Training Loss Plot

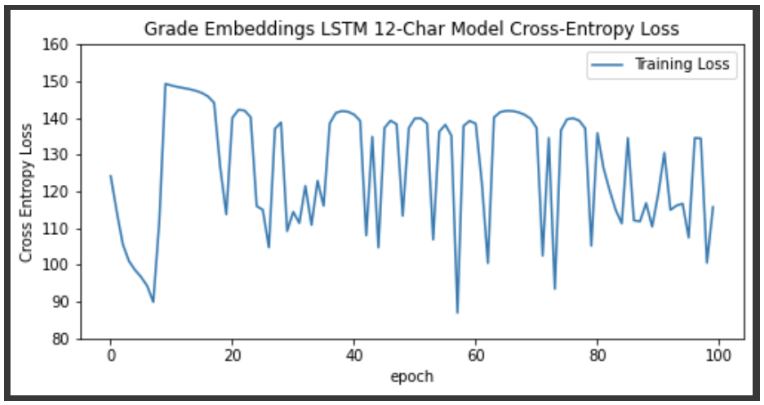

Figure 3: LSTM with Grade Embeddings Model Training Loss Plot

For the RoBERTa encoder, we fine-tuned our model on our data by using and modifying the Huggingface transformers run mlm.py and running the training process for 10,500 epochs. The hyperperameters in our setting were:

- hidden size: 768
- hidden droput probability: 0.1
- number of attention heads: 12
- number of hidden layers: 12
- vocabulary size: 50,265
- position embeddings: 514

After applying finetuning with the encoder model, we ran an evaluation on the validation data set. The plot displayed below (see Figure 4) shows the validation cross-entropy loss over time for this model setting.

5.4 Results

Table 1 contains the course-level accuracies and cross entropy losses for our three experiments.

As shown in Table 1, the Finetuned RoBERTa model performs significantly better than the simpler LSTM models. While this result is expected due to the high performing ability of a robust pre-trained

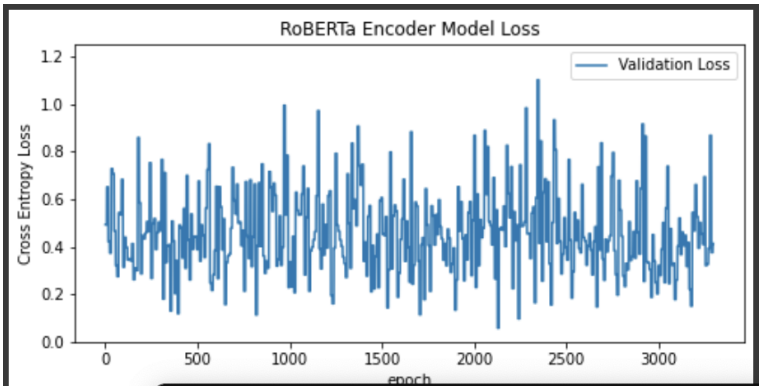

Figure 4: Fine-tuned Encoder Model (RoBERTa) Validation Loss Plot

Table 1: Accuracy and Cross-Entropy Loss Score for Our Models

Model	Accuracy	Cross Enropy Loss
LSTM Model with Course Sequence Embeddings (Baseline)	20%	95.0
LSTM Model with Course Sequence + Grade Embeddings	10%	140
Finetuned RoBERTa Model	**89.01%**	**0.4593**

encoder, we did not expect the character-level LSTMs to perform so poorly. We had expected that character-level information would help capture important nuances of course strings, such as the subject department and course level, and hence improve predictions on course sequences. These results may suggest a deeper bidirectional representation of course names is important for course prediction tasks, and that there are longer range dependencies between courses taken by students than we might expect.

While we wrote a custom character-matching evaluation method and used character-level BLEU scores, many predictions ended up having all characters that were completely different from course string labels. Thus, these metrics did not give us more representative results than the course-level accuracy that we had previously computed.

6 Analysis

As mentioned previously, there are reasons to expect that the Finetuned RoBERTa model would significantly outperform the simple LSTM model. However, it is interesting that the encoder model's pre-training on English language Masked Language Modeling tasks is helpful also for Masked Language Modeling tasks for course strings, even though course strings are unlike English words, in that they have different types of characters and combinations.

On probing the outputted cross-entropy losses (negative log-likelihoods) for each of the test examples for the Encoder model, we see the following:

The most unnatural (unlikely) course sequence - with the highest cross entropy loss - was:
ATHLETIC2,ATHLETIC134,CHEM33,HUMBIO4S,IHUM30B, PSYCH11N,CHEM33,
DANCE46,IHUM30A,PSYCH1,CHEM31,IHUM19,STS50A,WCT3B

The most natural (likely) course sequence - with the lowest cross entropy loss - was:
MATH51,ATHLETIC132,CHEM24N,POLISCI140,IHUM38B,ATHLETIC78,CHEM33,
SPANLANG22B,INTNLREL191,PWR1,IHUM38A,CHEM31,SPANLANG21B,INTNLREL191,
IHUM19,HISTORY53N

As we can see in the above sequences, the most natural sequence involves courses in more subjects (9) than that in the least natural sequence (5). Further, on examining the results from the fine-tuned encoder for each test example, we observed that it was more natural for students to take

courses in a larger number of distinct subjects than in a lower number of distinct subjects. While this could seem counterintuitive, it probably demonstrates the variation in courses being explored by freshmen, which is typically encouraged at liberal arts colleges such as Stanford.

The plot below displays the number of different subjects taken for test course sequences, shown in increasing order of cross-entropy loss.

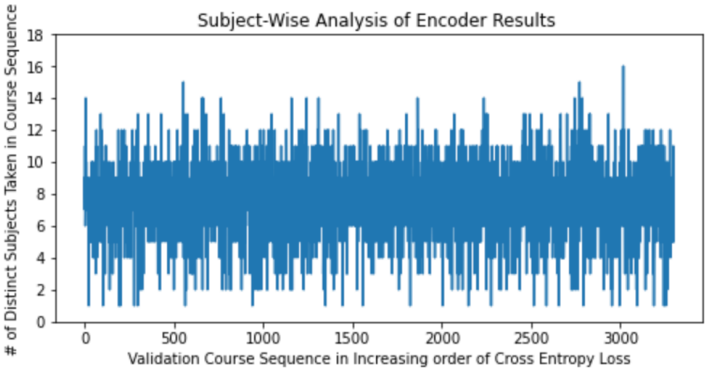

Figure 5: Number of Different Subjects Taken for Test Course Sequences with Increasing Cross Entropy Loss

7 Conclusion

The main takeaways from this work, are the following:

- A deep bidirectional pretrained encoder model performs much better on course prediction tasks than a character-level simple stacked LSTM model.
- It is more natural for students to take courses in a large number of distinct subjects in freshman year at Stanford than in a small number of distinct subjects.

Some limitations of our work are that the LSTM model outputs predictions that do not necessarily look like valid Stanford course strings. This does not allow us to make any claims about how students' grades impact course trajectories, as we had originally hoped to analyze. Time and resource constraints restricted our analysis to all freshman-year course trajectories at once rather than a subject-by-subject deep dive. It is likely that different degree plans vary in their relationship between the naturalness of course trajectories, number of distinct subjects taken, and persistence along a degree path. Future analyses will address this research space.

Future work can focus on analyzing and comparing the naturalness (likeliness) of course sequences for other academic years, and whether it is more likely for students to take courses in a large number of distinct subjects than a lower number of distinct subjects. Further, future work can also analyze the influence of grades on course trajectories.

References

[1] A. Elbadrawy and G. Karypis. Domain-aware grade prediction and top-n course recommendation. *RecSys '16.*, 2016.

[2] Gelbgiser D. Morgan S. L. Weeden, K. A. Pipeline dreams: Occupational plans and gender differences in stem major persistence and completion. *Sociology of Education, 93(4), 297–314.*, 2020.

[3] Pascarella E. T. Wolniak, G. C. The effects of college major and job field congruence on job satisfaction. *Journal of Vocational Behavior*, 2005.

[4] Degree planning with plan-bert: Multi-semester recommendation using future courses of interest. 35.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[7] David N Lang, Alex Wang, Nathan Dalal, Andreas Paepcke, and Mitchell Stevens. Forecasting undergraduate majors using academic transcript data, Nov 2021.

[8] W. Jiang and Z. A. Pardos. Time slice imputation for personalized goal-based recommendation in higher education.

[9] Ning X. Lan A. Ren, Z. and H. Rangwala. Grade prediction based on cumulative knowledge and co-taken courses. *In Proceedings of the 12th International Conference on Educational Data Mining. ERIC.*, 2019.