Prediction of Undergraduate Students' Course Sequences and their Naturalness

Shyamoli Sanghi, Victoria Delaney, Qi Han
{shyamoli, vdoch, qihan96}@stanford.edu,

Stanford
CS224N, Winter 2022

Objective and Background

Predicting undergraduates' future course trajectories given a sequence of early academic courses is a relevant problem that will help institutions as well as students gain a better sense of what course path to select. While major forecasting has been explored qualitatively and quantitatively through various modeling and prediction techniques ([1],[2]), there has been less work on sequence prediction of students' courses when grades are integrated with respect to students' future course selections.

Our work uses NLP models to (1) predict **future courses**, given sequences of past freshman year courses and (2) predict how natural a given course sequence is, as well as **how many distinct subjects are taken naturally**, in freshman year.

Problem Setup

1. We want to predict future course, given sequence of past freshman year courses (+sequence of past grades)
2. We also want 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 number of distinct subjects taken in freshman year.

Method

Feature Representations: We examined only freshman courses, represented by:

- Course Catalog + Course Number Representation, comma-separated
- One-hot encodings per character in the course sequence (39 total chars.)
- Padding on course sequences shorter than the max. sequence (252 chars.)
- One-hot encodings per grade type (49 total unique grades)

Models: (1) We created a custom stacked LSTM with course and grade embeddings and (2) We fine-tuned the encoder RoBERTa([3]) on our data.
LSTM Model Summary: We created a **character-level** sequence model that comprises 2 LSTM layers and one Linear layer. The first LSTM layer outputs a hidden state matrix with dimensions that represent (sequence length)*(1st hidden size) (with return sequences = True). The second LSTM's hidden state has dimension = smaller than the 1st hidden size (with return sequences = False).

Encoder Model Summary: We used RoBERTA[3] to fine-tune on our data, at a **course-level** (word level, not character level) . Since RoBERTa is trained on Masked Language Modelling and Next Sentence Prediction tasks, we use this model to determine the perplexity and the naturalness of a given course sequence.

Data and Pre-Processing

The data set records the anonymous enrollment data for 20 years (2000-2020) of 26,892 undergraduate students at Stanford University. Each row in the dataset corresponds to a (course, student) enrollment decision. The relevant variables from our dataset are (1) student identifier, (2) course enrollment term and year, (3, 4) course subject and catalog number, (5) student's final grade and (6) degree plan. We concatenated the course subject and catalog number to create full course names and extracted students' freshman course histories to create course sequences. An example course sequence is shown here: ["CS106a,CS221,AA228,EDUC424"]. Students' first sophomore year courses were extracted and served as our ground-truth course labels.

Experiments

Grade-Level Embeddings LSTM:

- ReLU activation functions between both hidden layers
- Adam optimizer, Cross Entropy Loss
- Batch size 128 (due to encodings' large size)

RoBERTa Encoder: fine-tuned our model with parameters relevant to the dataset, trained for 10,500 epochs.

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

Evaluation Metrics and Results

For the two LSTM experiments, we computed the following:
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 model output course sequences, where we took a course sequence to represent a sentence of characters.
For the encoder model, with 3,299 test examples, we obtained overall perplexity of 1.583. Below are the course-level accuracies and cross entropy losses for the three experimental settings.

Model	Accuracy	Cross Entropy Loss
Baseline LSTM Model	20%	95.0
LSTM Model + Grade Embeddings	10%	140.0
Finetuned RoBERTa Model	**89.01%**	**0.4593**

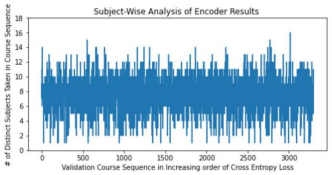
Analysis and Observations

The performance of the encoder RoBERTa is much higher than the LSTM model, which may suggest that:

1. 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.
2. Even though the former is a character-level model, which could capture more complex nuances of course names, the LSTM model is too simplistic compared to the complex pre-trained RoBERTa model.

Examining the results from the fine-tuned encoder for each test example, we observed that: the **least natura**l number of distinct subjects was **5** and the **most natural** number of distinct subjects taken was **9**. 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.



Conclusions

- 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 than in a small number of distinct subjects.

References

[1]Shao, Erzhuo, Shiyuan Guo, and Zachary A. Pardos. "Degree Planning with PLAN-BERT: Multi-Semester Recommendation Using Future Courses of Interest." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 17. 2021.
[2]David N Lang, Alex Wang, Nathan dalal, Andreas Paepcke, and Mitchell Stevens. Forecasting undergraduate majors using academic transcript data, Nov 2021
[3]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.