Degree Planning with PLAN-BERT: Multi-Semester Recommendation Using Future Courses of Interest

Erzhuo Shao[1], Shiyuan Guo[2], Zachary A. Pardos[3]*

[1] Department of Electronic Engineering, Tsinghua University
[2] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
[3] Graduate School of Education, University of California, Berkeley
sez20@mails.tsinghua.edu.cn, {shiyuan.guo, pardos}@berkeley.edu

Abstract

Planning scenarios involving user pre-specified items present themselves frequently in recommender system domains. Although next-item and next-basket recommendation has been a focus of prior research, multiple consecutive item or basket approaches are needed for planning. No prior work has leveraged pre-specified future reference items to improve this type of challenging consecutive prediction task at inference time. PLAN-BERT is the first to accommodate this general planning scenario. It does so by contributing novel modifications that take inspiration from the masked training and contextual embedding of self-attention models. To test the model, we use the domain of student academic degree planning, in which students' past course histories and future pre-specified courses of interest are used to fill in the remainder of their curriculum. Our offline analyses consist of 15 million historic course enrollments at 20 institutions and an online evaluation conducted at one of the institutions. Our results show that PLAN-BERT outperforms existing models including BERT, BiLSTM, and a UserKNN baseline, with small numbers of future reference items substantially improving accuracy. Significant results from our online evaluation show PLAN-BERT to be strongest in students' perceptions of personalization.

Introduction

Research on session-based predictive approaches have focused on next time slice recommendation (Fang et al. 2019; Mehta, Hofmann, and Nejdl 2007; Berkovsky, Kuflik, and Ricci 2007). Much less explored have been models that predict multiple consecutive items, or baskets, for recommendation (Cheng et al. 2013). Unexplored has been the use of future reference items to aid in this task. Consecutive time slice recommendations are called for in planning scenarios and it may be desired or even necessary for a user to select several individual (i.e., item) or grouped (i.e., basket) future items ahead of time. Multiple time slice recommendation is a challenging prediction problem as auto-regressive models build on earlier assumptions, amplifying error over time as a result (Liu et al. 2019c). Incorporation of user pre-specified future items into planning-based models can potentially mitigate this degradation in accuracy and improve the personalization of recommendations. We investigate this potential in both offline and online experiments with the introduction of PLAN-BERT, the first model to support pre-specified future reference items to make multiple consecutive recommendations without auto-regression.

We chose Bidirectional Encoder Representations from Transformer (BERT) (Devlin et al. 2019) as the base architecture for our approach due to the similarity of BERT's masked token prediction objective to our planning recommendation scenario. This objective randomly masks a percentage of tokens during training, with the objective of reducing error in predicting them based on context. This is somewhat analogous to planning-based scenarios in which we would like a recommender to utilize both users' past items and pre-specified future items to generate a plan. Unlike with BERT applied to natural language, where 15% of tokens are conventionally masked, the majority of tokens may be masked or missing in planning-based scenarios if a user's pre-specified future items represent only a small portion of an expected plan. We evaluate PLAN-BERT on consecutive basket recommendation, though consecutive and next-item recommendation is a supported special case. Our introduction of PLAN-BERT makes the following novel modeling contributions:

- Inference time utilization of future reference items for consecutive recommendation
- Explicit user and item features as input to BERT for recommendation
- Multiple consecutive basket prediction without auto-regression

We evaluate PLAN-BERT on the high stakes task of student academic degree planning. This task is well suited for planning-based recommendation, with semesters of enrollments akin to baskets of items, and consecutive semester recommendations expected to be made based on past course histories and future pre-specified courses of interest. Support for academic planning is important because of the stakes involved. Course selection decisions can greatly affect students' careers and enrollment is costly, with annual tuition ranging from thousands to tens of thousands of dollars. As noted by Elbadrawy and Karypis (2016), this is also a difficult recommendation task due to students' multifaceted selection criteria. The task is not only difficult for predictive models, but for humans too. Completion rates for students at

*Corresponding author.

4-year institutions hover at around 62% (Shapiro and Dundar 2016), with research suggesting that this is in part due to the multitude of decisions involved in degree planning (Baker 2018). With student-adviser ratios at large institutions around 1:600 (Carlstrom and Miller 2013), and tightening institutional budgets (Dowd and Shieh 2014; Mitchell, Leachman, and Saenz 2019), there is a national need for effective recommender systems development in this area.

Our offline analyses of a 20 institution dataset, consisting of both 2-year community colleges and 4-year degree granting institutions, show how PLAN-BERT outperforms BERT and other deep and non-deep model baselines and the benefit of future reference courses. We conduct an online user study, implementing PLAN-BERT and other models into a production course planning system. Statistically significant results from the study find that PLAN-BERT scored highest in student perceptions of personalization.

Related Work

We briefly review work in three areas closely related to ours: sequential next-item and next-basket prediction, approaches to the user-wise cold start problem in recommendation, and approaches to course recommendation in higher education.

Next-Item and Next-Basket Recommendation

Next-item sequential recommendation has been frequently studied in recommender systems (Yap, Li, and Philip 2012; Li, Zhao, and Liu 2018), where patterns in user history are used to predict the next item a user may access. More recently, efforts have explored next-basket recommendation, where a user might, for example, add multiple related items to a shopping cart (Rendle, Freudenthaler, and Schmidt-Thieme 2010; Wan et al. 2015; Yu et al. 2016). Deep model approaches using recurrent and convolutional neural nets have been most commonly used to make next-item, within or across session recommendations by modeling user sequential item or clickstream histories (Hidasi et al. 2015, 2016; Li et al. 2017; Quadrana et al. 2017; Tuan and Phuong 2017; Wang et al. 2018; Guo et al. 2020).

Recent advances in natural language processing have revealed that multi-head self-attention models have strong sequence modeling capabilities (Vaswani et al. 2017; Radford et al. 2019; Devlin et al. 2019). One model in particular, BERT (Devlin et al. 2019), has been used to great effect in NLP. BERT uses a bidirectional self-attention architecture that predicts randomly selected masked words in text during training. Sun et al. (2019) adapted this model to the next-item recommendation task by changing the training objective to predict masked items and attain state of the art performance on several recommendation datasets.

Approaches to the User-Wise Cold Start Problem

The user-wise cold start problem refers to when recommendation models have little to no data about a user with which to personalize recommendations (Ralph et al. 2019). Collaborative filtering-based approaches have used side information, such as user profile information and user social connections to assist in user-wise cold start recommendation (Sedhain et al. 2014; Barjasteh et al. 2015). Other work has taken into account user interface actions to infer user intent and thus improve cold start personalization (Wu et al. 2016). In a similar vein, Sun et al. (2013); Christakopoulou, Radlinski, and Hofmann (2016) selectively chose questions to ask users regarding their preferences to quickly profile users and demonstrated that answers to these questions improved recommendation performance over baselines. Recommending courses to new freshmen with no course history or declared major is an example of the user-wise cold start problem in higher education. Asking students for their intended major is one approach to overcoming this problem (Pardos, Fan, and Jiang 2019). Asking students to specify courses they would like to take in future semesters is another approach, which we explore with PLAN-BERT.

Course Recommendation

Several facets of course recommendation have been explored in past approaches. Non model-based deployed systems display analytics to students drawn from aggregate course evaluations (Chaturapruek et al. 2018). Other recommender systems have focused on degree requirements and constraint satisfaction as priorities for recommendation (Parameswaran, Venetis, and Garcia-Molina 2011) and the scheduling interfaces for accommodating these constraints (Li, Tinapple, and Sundaram 2012). Predictive models have been employed for next course recommendation, utilizing student ratings of courses (Farzan and Brusilovsky 2006; Bendakir and Aïmeur 2006) and student and course hierarchical features (Elbadrawy and Karypis 2016).

Neural approaches to course recommendation have begun to emerge, utilizing sequential enrollment histories for next-course enrollment (Pardos, Fan, and Jiang 2019; Polyzou, Athanasios, and Karypis 2019), grade prediction (Ren et al. 2019; Jiang, Pardos, and Wei 2019), and course similarity prediction (Pardos and Nam 2020; Pardos and Jiang 2020). Only a constraint-based approach (Parameswaran, Venetis, and Garcia-Molina 2011) has focused on long-term planning, by constraining the recommendation space of existing recommender scores using student graduation and course requirements. Recent work has addressed short-term planning for a desired next semester course outcome (Jiang and Pardos 2019). This RNN-based approach supports only single time slice recommendation due to the exponential complexity of considering different sets of course inputs across multiple time slices. PLAN-BERT addresses this single-basket limitation, but focuses on incorporating future courses of interest to generate a long-term coherent plan, rather than explicitly optimizing for a single future course outcome goal.

PLAN-BERT

To address the modeling challenges of consecutive basket prediction for planning-based recommendation scenarios, we adapt a Transformer-based model to generate degree plans; we designate this adaptation PLAN-BERT. In this section, we will describe the incorporation of future reference courses and historical courses into the BERT framework. We will then introduce user and item features into the

model, overview the architecture of PLAN-BERT, and then describe the training procedure.

Prediction with Past and Future Reference Courses

In the scenario of higher education, similar to next-basket recommendation, items (i.e., courses) in the same basket (i.e., semester) occur in parallel. Therefore we employ a sequence of multi-hot vectors to represent the course schedule of a student, denoted by $C^{t\times\|\mathbb{C}\|} \in \{0,1\}$, where t is the number of semesters and $\mathbb{C}$ denotes the set of all courses at the institution. We partition a student's t semesters into historical and future semesters. We denote the number of historical semesters as h, and define the historical course sequence of a student to be $H^{h\times\|\mathbb{C}\|} = C[0:h]$.

Students often have several courses in mind that they are interested in taking in the future. They may be courses they plan to take with friends, courses related to a tangential interest, or upper level courses of contemporary relevance to their intended research or industry careers. To model students' multifaceted preferences, we utilize these future courses of interest r as input to PLAN-BERT. Incorporation of these courses also allows for personalized plans to be generated even for new freshmen with no course history and no declared major. By exploring the impact of the number of future reference courses specified, r, on accuracy, we can attempt to identify the lowest number that can overcome the cold start problem without presenting an undue burden for the user to provide more information than necessary before generating their plan. Providing pre-specified future reference courses to PLAN-BERT can also help reduce prediction error across long sequences of baskets, as the reference courses serve as anchors for the output sequence space.

For our offline analyses, we randomly sample r courses from the complete course schedule C to serve as example references courses. We accept as a limitation that the actual courses students are interested in may be of a different distribution from this sample. The sampled referenced courses are represented by $R^{t\times\|\mathbb{C}\|} = Sample(C,r)$ and we have $\sum R^{t\times\|\mathbb{C}\|} = r$. In our scenario, students preselect their reference courses when they enter the university. Thus, we employ $K = [C[0:h]; Sample(C,r)[h:t]]$ to represent all known courses, which is the concatenation of historical courses sequence H and future reference courses taken from the last $t-h$ rows of R. An example of the input and prediction target of a sophomore is demonstrated on the right of Figure 1, where the vertical dimension denotes the space of courses and the horizontal dimension denotes semesters. Blue, green, and yellow squares represent historical courses, $H = C[0:h]$, future reference courses, $R[h:t] = sample(C,r)[h:t]$, and target courses, $T = C - [C[0:h]; sample(C,r)[h:t]]$, which are masked in the input. The input of PLAN-BERT consists of blue and green courses while yellow target courses are expected output. We note that historical and reference courses are excluded from the prediction target.

Incorporation of Explicit User and Item Features

Prior work in NLP has leveraged features of words (Liu et al. 2019a) and document meta-data (Ostendorff et al. 2019) to improve text classification accuracy with BERT. We draw from these works and incorporate meta-data of users and items into PLAN-BERT.

In higher education, many auxiliary features of students and courses are available. *User features*, such as student major, could help reveal students' preferences and assist with the cold-start problem for new freshmen. *Item features*, such as course department, could similarly help approximate the embedding of a course that has only been offered for one semester.

We use matrix $U^{t\times\|\mathbb{U}\|}$ to represent one type of feature of a student, where $\|\mathbb{U}\|$ is used to denote the size of the vocabulary of a user feature and t denotes the number of semesters of the student. For each student with h historical semesters, we know her user features of the first $h+1$ semesters. We therefore concatenate the user features of the first $h+1$ semesters and zero pad the remaining $t-h-1$ semesters, formulated as $F_u = [U[h+1:t], 0^{(t-h-1)\times\|\mathbb{U}\|}]$. Similarly, we represent item features as $I^{\|\mathbb{C}\|\times\|\mathbb{I}\|}$, where $\|\mathbb{I}\|$ denotes the size of the vocabulary of an item (course) feature and $\|\mathbb{C}\|$ denotes the number of all offered courses. Since only the features of known courses are available for recommendation, we employ the product of known courses and item features' matrix, $F_i = K^{t\times\|\mathbb{C}\|} I^{\|\mathbb{C}\|\times\|\mathbb{I}\|}$ as the representation of item features.

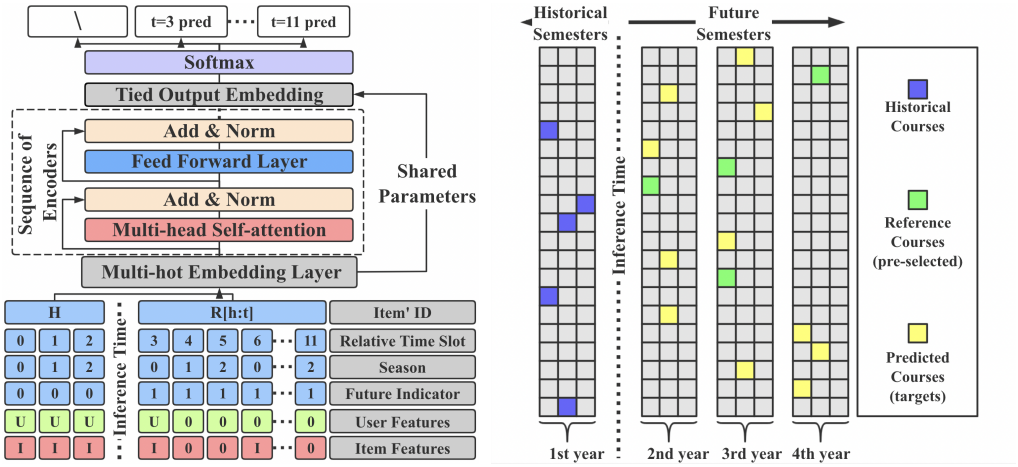
In addition to user and item features, there also exist features that are the result of an interaction between both users and items. One example is course grades earned by students, thought to play a factor in course selection. However, prior work on predicting grades has proven this to be a difficult task for neural models (Jiang, Pardos, and Wei 2019; Ren et al. 2019), perhaps owing to their non-normal distribution (Arthurs et al. 2019). Furthermore, inputting grades has been found not to improve next semester course enrollment prediction (Pardos, Fan, and Jiang 2019). We therefore leave the exploration of methods for meaningfully incorporating course grades as a matter for future work.

Architecture

In this section, we introduce the architecture of PLAN-BERT. Fig. 1 illustrates (a) the components of this architecture and (b) the inputs and outputs of the model.

PLAN-BERT accepts several inputs; each represented by a $t \times x$ matrix, where t is the number of semesters and each row denotes the information of a corresponding semester. The inputs are as follows:

- **Historical Courses and Reference Courses:** A matrix of known courses represented by $K = [C[0:h]; Sample(C,r)[h:t]]^{t\times\|\mathbb{C}\|}$, which is the concatenation of historical semesters and future semesters containing reference courses.
- **Relative Semesters:** An identity matrix of size $t \times t$. It is the number of elapsed semesters since the student began and is equivalent to the positional encoding in



(a) The architecture of PLAN-BERT (b) The input and prediction targets of PLAN-BERT

Figure 1: (a) The architecture of PLAN-BERT and (b) the historical and future reference courses and prediction targets

BERT which provides positional information to the self-attention mechanism.

- **Semesters' Seasons:** A matrix denoted by $S_2^{t\times 3}$. Each dimension is a one-hot vector. It denotes the season (e.g., Fall, Spring, or Summer) of the corresponding semester.
- **Future Indicator:** Resembling the [mask] token in BERT, we introduce a vector with shape $T^{t\times 1} \in \{0, 1\}$ to indicate the time slots to be predicted. It is a list of tokens used to indicate future semesters. If the corresponding semester is future, the value is 1, otherwise it is 0.
- **User & Item Features:** Matrices for student $F_u = [U[h+1:t], 0^{(t-h-1)\times\|\mathbb{U}\|}]$ and course $F_i = K^{t\times\|\mathbb{C}\|}I^{\|\mathbb{C}\|\times\|\mathbb{I}\|}$ features as described in the previous subsection.

Based on the aforementioned inputs, the architecture of Devlin et al. (2019) is employed to generate degree plans. Devlin et al. (2019) utilize embedding layers for projecting all inputs into dense vector sequences and a series of bidirectional transformer encoder layers to produce contextual embeddings. Finally, a tied embedding layer (Inan, Khosravi, and Socher 2017) and a softmax layer are used to project contextual embeddings back to the space of courses as a recommended course schedule.

Training Procedure

Inspired by masked language models like BERT, we employ percentage sampling to pre-train PLAN-BERT to learn the contextual embedding of courses. In this pre-training stage, we randomly select a proportion α of courses from all courses $C^{T,\|\mathbb{C}\|}$ as reference courses; PLAN-BERT does not utilize historical courses in pre-training, so $h = 0$ in this phase. The remaining $1-\alpha$ proportion of courses are the expected targets. The matrix of known courses in this phase is $K = Sample(C, \alpha\sum C)$ and the target is $C - K$.

PLAN-BERT learns the relationship between history and future reference courses by utilizing a fine-tuning stage where we randomly sample r courses from $C^{T,\|\mathbb{C}\|}$ as reference courses R and we randomly select an integer h among $0, 1\cdots T-1$ as the number of historical semesters. Therefore, we have a number of historical semesters $h = random(0, T-1)$, known courses $K = [C[0:h]; sample(C, r)[h:t]]$, and target of output $C - K$ in this phase. In both pre-training and fine-tuning, to reduce overfitting, PLAN-BERT samples reference courses and the number of historical semesters h dynamically, which means we change their random seeds in each epoch as was done in Liu et al. (2019b).

Offline Evaluation

We firstly describe the datasets and experiment settings for offline evaluation of PLAN-BERT and other comparison models. We then report results of our offline analyses comparing PLAN-BERT to baseline models on the task of multi-semester enrollment prediction, then break-down PLAN-BERT predictive performance by student year, semester, and number of future reference courses used.

Datasets

We use an anonymized dataset of undergraduate enrollment from the University of California at Berkeley, a pubic liberal arts university in the USA (referred to as UNIVERSITY1), and a large dataset from the City University of New York, a system of 19 public colleges offering a mixture of associate's and bachelor's degrees, also in the USA (referred to as SYSTEM1). We use data from 17 of these colleges, since two colleges lacked sufficient data to evaluate model performance. We use the first two years of enrollment data from all colleges in SYSTEM1 to ensure all comparisons for SYSTEM1 use the same number of training and comparison semesters. The UNIVERSITY1 dataset consists of 4.6 million course enrollments in 7,252 unique courses by 134,275 bachelor's degree seeking students between Fall 2008-2017. Student features in this data consist of major, the college, department, and division of the major,

Hyperparam.	Searched	Val. (U1)	Val. (S1)
Layers	$[1, 4]$ by 1	2	3
Hidden Size	$128, 256, 512, 1024$	512	512
Dropout	$[0.0, 0.5]$ by 0.1	0.2	0.2
α (1 - mask%)	$[0.4, 0.9]$ by 0.1	0.8	0.8
Finetuning r	$[3, 15]$ by 1	10	5
Conf. Pen. β	-	0.1	0.1
Learn Rate	-	10^{-4}	10^{-4}

Table 1: Hyperparameter search ranges and values used in the evaluation of UNIVERSITY1(U1) and SYSTEM1(S1).

and whether the student entered as a transfer or new freshman. Course features consist of the department, subject, and instructors of each course. The SYSTEM1 dataset consists of 10.4 million course enrollments made by 903,226 total students across the 19 institutions between Fall 2014-2019. The median number of courses per institution was 1,177 and the median number of students per institution was 45,287. The student and course features of SYSTEM1 were similar to that of UNIVERSITY1.

These datasets have not been made publicly available, as the possibility of re-identification is significant and would violate federal protections against disclosure of student records (FERPA 1974). Access to these data may be obtained by establishing a data access agreement with the Office of the Registrar responsible for each dataset.

Experiment Settings

For both datasets, we partition students temporally by starting year to create the testing sets. The last four years of enrollment data comprise the test set of UNIVERSITY1 and the last two years for SYSTEM1. We generate the validation and training sets as a 20/80 random split of the students not included in the test set. We note that in both datasets, there are new courses offered in the testing period which have never been seen in the training period, which therefore cannot appear in the predicted course plans and lower our recall. In UNIVERSITY1, 24.8% of courses in the testing set were new, and 12.84% were new, on average, in SYSTEM1. Finally, we employ $r = 5$ reference courses for each student, derived from our experiments in the Impact of Reference Courses section of the results.

We repeat all experiments five times and report average results, setting random seeds per experiment. We report results trained with the best hyperparameters found using a grid search, shown in Table 1. We follow (Pardos, Fan, and Jiang 2019) in using Recall@10 as the primary evaluation metric in experiment performance evaluations. We also present overall summary evaluation results using the popular NDCG@10 metric. Recall@10 is $\frac{|y \cup \hat{y}[:,0:10]|}{|y|}$, or the proportion of actual courses taken contained in the top 10 predicted courses taken in each semester, where y denotes the actual courses taken and $\hat{y}$ denotes predicted courses taken. Normalized Discounted Cumulative Gain (NDCG) is a commonly used evaluation metric that evaluates the degree of relevance of each item (Järvelin and Kekäläinen 2002).

For each model, we stop its pre-training and fine-tuning when validation Recall@10 has not increased for 10 epochs. Experiments concluded after one week on a system with 4x NVIDIA GTX 980Ti, 2x Xeon E5-2620 v3 CPU, and 256GB RAM using Python 3.6.3 and Keras 2.3.0 on Ubuntu 18.04. Model and experiment code is available online[1].

Baselines

We compare three baselines and user and item variations on our proposed PLAN-BERT model. These baselines were chosen because of their competitiveness and ease of adaptation to the sequential basket recommendation problem. Because our task requires the prediction of items, the time slots of those items, and the incorporation of future reference items, adapting other classical and deep baselines such as $P^3\alpha$, ItemKNN (Dacrema, Cremonesi, and Jannach 2019), and S3-Rec (Zhou et al. 2020) is a non-trivial task that would merit extensive modification to the respective baseline architecture.

- **UserKNN (Sarwar et al. 2001):** A classic recommendation approach based on k-nearest-neighbors and user-user similarities which was found to be competitive against state of the art deep learning methods (Dacrema, Cremonesi, and Jannach 2019). For each user, we employ cosine similarities of known courses of the user and complete course schedules of historical students. We treat the entire history and reference courses of each student as a single basket rather than a sequence of baskets. We use user-user cosine similarities to calculate a weighted average of historical students' schedules.
- **LSTM (Hochreiter and Schmidhuber 1997):** A classical recurrent neural network model, which is widely used in sequential recommendation. In our paper, LSTM replaces Transformer encoder layers of PLAN-BERT with a sequence of LSTM layers. Since LSTM can only model historical information, only historical courses are provided and reference courses are excluded.
- **BiLSTM (Schuster and Paliwal 1997):** A bidirectional LSTM, modeling sequences in both forward and backward directions. Similar to LSTM, we directly replace the Transformer encoder layers of PLAN-BERT with BiLSTM layers. Reference courses are included in its input.
- **BERT (Devlin et al. 2019):** This represents the exact implementation of Devlin et al. (2019), in which future reference items and meta-features are not utilized. This represents BERT without our PLAN-BERT enhancements. We omit the next-sentence prediction loss due to the recommendation context, akin to BERT4Rec (Sun et al. 2019).
- **PLAN-BERT:** Our proposed model, whose training procedure includes pre-training and fine-tuning, with future reference courses utilized in training and inference. The **+user** and **+item** suffixes mean student features and course features have been added to the model in training and inference. PLAN-BERT with both of these additions represents our fully realized model.

[1] https://github.com/CAHLR/plan-bert-aaai

	Recall@10						NDCG@10					
Datasets	UNIVERSITY1				SYSTEM1		UNIVERSITY1				SYSTEM1	
Class standing	Fr	So	Jr	Sr	Fr	So	Fr	So	Jr	Sr	Fr	So
UserKNN	0.2317	0.1900	0.1764	0.1920	0.2822	0.2748	0.1779	0.1381	0.1325	0.1456	0.1737	0.2097
LSTM	0.1299	0.2158	0.2327	0.2388	0.1699	0.3523	0.0953	0.1618	0.1767	0.1778	0.1276	0.2631
BiLSTM	0.2555	0.2702	0.2787	0.2843	0.3622	0.3803	0.1945	0.1927	0.1875	0.1860	0.2202	0.2632
BERT	0.1778	0.2387	0.2881	0.2981	0.1662	0.3636	0.1333	0.1719	0.1908	0.2108	0.1241	0.2942
PLAN-BERT	0.2648	0.2851	0.2985	0.3225	0.3726	0.4222	0.2142	0.2191	0.2292	0.2469	0.2988	0.3555
PLAN-BERTi	0.2755	0.2958	0.3047	0.3275	0.3915	**0.4547**	0.2219	0.2265	0.2284	0.2401	0.3119	0.3706
PLAN-BERTu	0.2704	0.2867	0.3118	0.3327	0.3753	0.4241	0.2108	0.2095	0.2245	0.2386	0.3018	0.3643
PLAN-BERTiu	**0.2859**	**0.2967**	**0.3161**	**0.3338**	**0.3945**	0.4471	**0.2270**	**0.2281**	**0.2369**	**0.2510**	**0.3196**	**0.3722**

Table 2: Results of PLAN-BERT and baselines on UNIVERSITY1 and SYSTEM1 datasets.

Results

We examine enrollment plan generation performance of PLAN-BERT and comparison models on our two datasets. Recall@10 and NDCG@10 of enrollment predictions of all test set semesters by all models are reported in Table 2, aggregated by the class standing of the predicted students as of their first semester in the test set. As discussed in the introduction of our datasets, SYSTEM1 has only Freshman and Sophomore students because it consists of majority two-year degree granting institutions. The Freshman result for UNIVERSITY1, for example, is the average Recall@10 of each semester in the four years of predicted enrollments (i.e., generated 4-year plan) starting from the first semester of the UNIVERSITY1 test set. The Freshman result is an example of the cold-start scenario, as no course histories exist for these students at the time of plan generation.

We find that PLAN-BERT+user+item attains the best recall across almost all class standings of students and the best NDCG across all class standings in UNIVERSITY1 and SYSTEM1. We note that even without user+item features, PLAN-BERT outperforms all other models for all student classes. Compared to BiLSTM, another neural architecture that utilizes future reference courses, PLAN-BERT's margin of improvement increases with standings representing longer course histories, showing the advantage of the self-attention and contextual embedding architecture for making use of sequence histories. We also find that for upperclassmen, the addition of both item and user features provides a substantial boost in performance over the addition of user or item features alone. These features, which include major, department of major, and department of courses, perhaps become more important for prediction in a student's later years as they take more courses within the department of their major. Finally, we observe that the models that do not utilize future reference courses (i.e., LSTM and BERT) perform substantially worse on the cold-start scenario of predicting future enrollments for Freshmen.

We further investigate the performance of PLAN-BERT, breaking out the results of Table 2 by each semester predicted for each class standing. Figure 2 shows Recall@10 per semester for the PLAN-BERT+user+item model on the UNIVERSITY1 dataset. With 3, 6, and 9 historical semesters, Sophomore, Junior, and Senior students' Recall@10 reaches 0.2247, 0.2680, and 0.3592, improving by 12.07%, 30.07%, and 79.15% respectively over the Freshman recall at the same semester, showing that Recall@10 improves exponentially with additional historical semesters. Although many random sampling procedures are employed in the training and inference stages in our BERT-based models, their evaluation results are relatively stable: the maximum standard deviations of Recall@10 of BERT, PLAN-BERT, PLAN-BERT+item, PLAN-BERT+user, and PLAN-BERT+user+item across the five experiment repetitions for each class standing are 0.0046, 0.0017, 0.0022, 0.0021, and 0.0014, respectively.

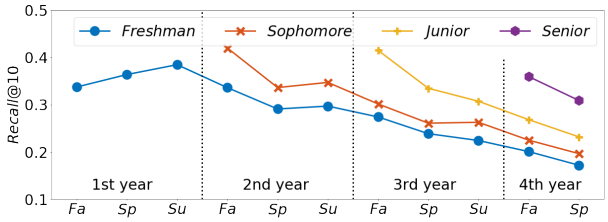
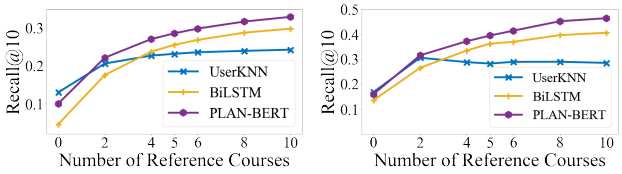


Figure 2: Impact of historical semesters on enrollment prediction on UNIVERSITY1 dataset, broken out by (Fa)ll, (Sp)ring, and (Su)mmer semesters.

Impact of Reference Courses The previous analysis observed the benefit of various amounts of historical data in predicting enrollments. In this section, we analyze the benefit of various amounts of future data (i.e., reference courses). Previous analyses used a fixed number of future reference courses ($r = 5$). In this analysis, we vary r from zero to ten and observe the impact on recall of the models which support future reference courses; UserKNN, BiLSTM, and PLAN-BERT variants. Figure 3 shows the results of plan prediction for Freshmen in both datasets. We observe that BiLSTM benefits the most from increased reference courses, exhibiting the steepest curve in UNIVERSITY1, while PLAN-BERT benefits most in SYSTEM1. The benefit of reference courses to UserKNN levels off at $r = 2$ in UNIVERSITY1 and begins to decrease after 2 in SYSTEM1. In both datasets, PLAN-BERT outperforms all models for every positive value of r.

Our offline analyses of future reference courses sample actual future courses from a student's enrollment sequence. However, in a real world scenario, which we evaluate in the



(a) UNIVERSITY1 results (b) SYSTEM1 results

Figure 3: Comparison of impact of r on recall for predictions of all semesters with no course history (i.e., Freshmen).

next section, these future courses of interest would need to be provided by the student. Users are often ambivalent to provide an excess of information upfront in order to receive a digital service in general (Chang, Harper, and Terveen 2015). We therefore would like to request as few reference courses as possible while still being able to effectively personalize a plan. The need for future reference courses will be greater for underclassmen, where the model has fewer historical courses to incorporate, or none at all, and students have the most semesters ahead of them. Figure 4 breaks down the impact of r on plan generation performance for each class standing using PLAN-BERT+user+item. Freshman and Sophomore plans are most affected by increased future reference courses. We also find that recall for Freshmen, who have no course history, improves 120% (0.1002 to 0.2214), from $r = 0$ to just $r = 2$ reference courses (Fig. 4), and improves over 190% (0.1002 to 0.2985), from $r = 0$ to $r = 6$ reference courses. We observe that the performance of $r = 2$ future reference courses for Freshmen is equivalent to Sophomores ($\sim$ 7 historical courses) with $r = 0$. We hypothesize that pre-specified future tokens provide a more informative bias towards the student's future intentions despite their sparsity compared to course history. Our results show that sparse pre-specified future reference items can greatly improve plan prediction performance and that PLAN-BERT's architecture is best suited, out of all models we considered in experimentation, to taking advantage of the information provided by pre-specified reference items.

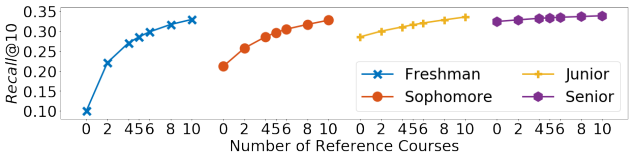


Figure 4: The recall of students of different grades with different r for PLAN-BERT+user+item in UNIVERSITY1 dataset. x-axis denotes r and y-axis denotes Recall@10.

Online Evaluation

We conduct an online user study by implementing our models into an existing production course recommender system[2] to explore if our offline evaluation agrees with subjective user-perceived performance of PLAN-BERT. We recruited 63 first and second-year undergraduate student participants from UNIVERSITY1 for our study. We filtered out 2 respondents that did not answer open-ended response questions according to the study directions. There was a wide representation of majors, with no one major representing more than 15% of respondents.

Evaluation Methodology

We follow the evaluation methodology of Adiwardana et al. (2020) for chat-bot model evaluation. They divide subjective evaluation into two phases: an interactive phase where raters converse with a chat-bot and rate their conversation, and a static phase where raters rate a set of completed conversations between a human and chat-bot. Adiwardana et al. (2020) adopt two rating metrics: sensibility, whether a conversation makes sense, and specificity, whether the chat-bot's responses are relevant to the topic being discussed. We adapted this methodology and defined sensibility to students as "whether the plan presented to [them] makes sense", and specificity as "how specific the plan is to [them]."

We use PLAN-BERT+user+item, LSTM, and UserKNN, as detailed in the previous baselines subsection to generate plans with the exception that the LSTM used for comparison in this online evaluation was the default model used in the production system, trained using an auto-regressive next-semester prediction approach following Pardos, Fan, and Jiang (2019). We limit the number of total plan generation models to ensure students had sufficient time to complete the study and randomize the order of models presented per respondent. We select UserKNN as a simple baseline, LSTM as the currently deployed production baseline, and PLAN-BERT+user+item as the top model from offline evaluations. We also add a randomly selected actual past student history of the same major as the respondent (referred to as "Actual") as a human-level plan generation benchmark for comparison.

Study Design

We split the survey of the study into three phases. In the first phase, *future course selection*, we ask students to select at least three courses of interest using a search tool to add those courses to their academic plan. Three reference courses struck a balance between attaining high recall performance in our offline analyses and minimizing the burden on students to add a large number of courses.

In the second phase, *interactive plan rating*, we present students with a set of three academic plans, generated dynamically from the models based on the respondent's course history and reference courses. These plans are presented in a randomized order and the student is asked to rate each plan in terms of its specificity and sensibility.

In the final phase, *static plan rating*, we randomly sample a past graduated student's history of the same major as the respondent and use this history as input to each model with $h = 2, r = 3$. We show respondents the past student's first two semesters and present the $r = 3$ as the courses the "example student wished to take in the future" and ask respondents to rate the plans generated for the next two years

[2]https://askoski.berkeley.edu

from each model. For comparison, we present the past student's actual courses taken as a candidate plan.

	Spec.	Sens.	Spec. (St)	Sens. (St)
PLAN-BERT	**72.13%**	54.10%	**77.05%**	47.54%
LSTM	65.57%	**55.74%**	70.49%	52.46%
UserKNN	65.57%	44.26%	60.66%	45.90%
Actual	-	-	**77.05%**	**60.66%**

Table 3: Plan ratings from the online user study. Spec. denotes specificity, sens. denotes sensibility, and (St) denotes ratings from the static plan phase.

Results

We report the percentage of respondents that answered "Yes" to whether a plan was sensible or specific in Table 3. We use a two-sided Wald test to compute p-values as in Winecoff et al. (2019). We compare among all valid respondents and apply a Bonferroni correction to account for multiple post-hoc comparisons.

Out of all models, PLAN-BERT is the only one to achieve statistically significant separation, being scored higher than UserKNN in specificity (p=0.0023) for ratings of other students' generated plans (St.) and attaining equal specificity to the actual student history. Specificity for actual student histories in the static rating context is also statistically significantly better than UserKNN (p=0.0023). These are the only statistically significant results, possibly owing to an insufficient sample size of only 63 students. Given sufficient statistical power, we would expect to see a difference between actual course plans and LSTM generated plans in specificity and sensibility. PLAN-BERT also achieves superior specificity in all contexts. Although LSTM attains higher sensibility in one context, actual student histories received an average sensibility of 60.66%, indicating it is quite difficult for students to evaluate sensibility. The results are similar to offline evaluation, where PLAN-BERT showed superior recall over LSTM and UserKNN.

Conclusion

We introduced PLAN-BERT, the first model to support consecutive basket recommendation with future reference items. Using a novel dataset of 20 institutions, we empirically demonstrated the value of pre-specified reference items in helping overcome the cold start problem. Even small numbers of pre-specified reference courses had a sizable impact, with $r = 2$ increasing Recall@10 of generated course plans by 120%.

Our results also demonstrate that PLAN-BERT's bidirectional self-attention architecture is better suited to utilize past sequence information than BiLSTM and a UserKNN baseline, and that incorporation of user and item features provided substantial benefit to Recall@10 in a student's later years of study. Finally, our online study showed that PLAN-BERT has plausible real-world applicability in providing virtual guidance to students navigating the complex terrain of a higher education degree. Future work may explore the application of PLAN-BERT to other domains such as itinerary planning or e-commerce recommendation.

Acknowledgements

We thank the Office of Institutional Research & Assessment at the City University of New York and the Office of the Registrar and Enterprise Data & Analytics at UC Berkeley for their anonymized data provisioning. This research was supported by grants from Schmidt Futures and Ithaka S+R.

References

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* .

Arthurs, N.; Stenhaug, B.; Karayev, S.; and Piech, C. 2019. Grades Are Not Normal: Improving Exam Score Models Using the Logit-Normal Distribution. *Proceedings of the 12th International Conference on Educational Data Mining* 252–257.

Baker, R. 2018. Understanding college students' major choices using social network analysis. *Research in higher education* 59(2): 198–225.

Barjasteh, I.; Forsati, R.; Masrour, F.; Esfahanian, A.-H.; and Radha, H. 2015. Cold-start item and user recommendation with decoupled completion and transduction. In *Proceedings of the 9th ACM Conference on Recommender Systems*, 91–98.

Bendakir, N.; and Aïmeur, E. 2006. Using association rules for course recommendation. In *Proceedings of the AAAI Workshop on Educational Data Mining*, volume 3, 1–10.

Berkovsky, S.; Kuflik, T.; and Ricci, F. 2007. Distributed collaborative filtering with domain specialization. In *Proceedings of the 2007 ACM conference on Recommender systems*, 33–40.

Carlstrom, A.; and Miller, M. A. 2013. National Survey of Academic Advising. URL https://www.nacada.ksu.edu/Portals/0/Clearinghouse/documents/Chapter6-ProfessionalAdvisorLoad-FINAL.pdf.

Chang, S.; Harper, F. M.; and Terveen, L. 2015. Using groups of items for preference elicitation in recommender systems. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1258–1269.

Chaturapruek, S.; Dee, T. S.; Johari, R.; Kizilcec, R. F.; and Stevens, M. L. 2018. How a data-driven course planning tool affects college students' GPA: evidence from two field experiments. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 1–10.

Cheng, C.; Yang, H.; Lyu, M. R.; and King, I. 2013. Where You like to Go next: Successive Point-of-Interest Recommendation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*.

Christakopoulou, K.; Radlinski, F.; and Hofmann, K. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 815–824.

Dacrema, M. F.; Cremonesi, P.; and Jannach, D. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, 101–109.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Dowd, A. C.; and Shieh, L. T. 2014. The Implications of State Fiscal Policies for Community Colleges. *New Directions for Community Colleges* URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cc.20120.

Elbadrawy, A.; and Karypis, G. 2016. Domain-aware grade prediction and top-n course recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 183–190.

Fang, H.; Zhang, D.; Shu, Y.; and Guo, G. 2019. Deep Learning for Sequential Recommendation: Algorithms, Influential Factors, and Evaluations. *arXiv: Information Retrieval* URL https://arxiv.org/abs/1905.01997.

Farzan, R.; and Brusilovsky, P. 2006. Social navigation support in a course recommendation system. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, 91–100. Springer.

FERPA. 1974. Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99). URL https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

Guo, Q.; Sun, Z.; Zhang, J.; and Theng, Y.-L. 2020. An attentional recurrent neural network for personalized next location recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 83–90.

Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; and Tikk, D. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* URL https://arxiv.org/abs/1511.06939.

Hidasi, B.; Quadrana, M.; Karatzoglou, A.; and Tikk, D. 2016. Parallel recurrent neural network architectures for feature-rich session-based recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 241–248.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-term Memory. *Neural computation* 9: 1735–80. doi:10.1162/neco.1997.9.8.1735.

Inan, H.; Khosravi, K.; and Socher, R. 2017. Tying Word Vectors and Word Classifiers: A Loss Framework for Language Modeling. *ArXiv* abs/1611.01462.

Järvelin, K.; and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20: 422–446.

Jiang, W.; and Pardos, Z. A. 2019. Time slice imputation for personalized goal-based recommendation in higher education. In *Proceedings of the 13th ACM Conference on Recommender Systems*, 506–510.

Jiang, W.; Pardos, Z. A.; and Wei, Q. 2019. Goal-based course recommendation. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 36–45.

Li, J.; Ren, P.; Chen, Z.; Ren, Z.; Lian, T.; and Ma, J. 2017. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1419–1428.

Li, Z.; Tinapple, D.; and Sundaram, H. 2012. Visual planner: beyond prerequisites, designing an interactive course planner for a 21st century flexible curriculum. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, 1613–1618.

Li, Z.; Zhao, H.; and Liu, Q. 2018. Learning from History and Present: Next-Item Recommendation via Discriminatively Exploiting User Behaviors. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; and Wang, P. 2019a. K-BERT: Enabling Language Representation with Knowledge Graph. *ArXiv* abs/1909.07606.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* abs/1907.11692.

Liu, Y.; Yu, R.; Zheng, S.; Zhan, E.; and Yue, Y. 2019c. NAOMI: Non-Autoregressive Multiresolution Sequence Imputation. In *Advances in Neural Information Processing Systems 32*.

Mehta, B.; Hofmann, T.; and Nejdl, W. 2007. Robust collaborative filtering. In *Proceedings of the 2007 ACM conference on Recommender systems*, 49–56.

Mitchell, M.; Leachman, M.; and Saenz, M. 2019. State Higher Education Funding Cuts Have Pushed Costs to Students, Worsened Inequality. URL https://www.cbpp.org/sites/default/files/atoms/files/10-24-19sfp.pdf.

Ostendorff, M.; Bourgonje, P.; Berger, M.; Schneider, J. M.; Rehm, G.; and Gipp, B. 2019. Enriching BERT with Knowledge Graph Embeddings for Document Classification. *ArXiv* abs/1909.08402.

Parameswaran, A.; Venetis, P.; and Garcia-Molina, H. 2011. Recommendation systems with complex constraints: A course recommendation perspective. *ACM Transactions on Information Systems (TOIS)* 29(4): 1–33.

Pardos, Z. A.; Fan, Z.; and Jiang, W. 2019. Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance. *User Modeling and User-Adapted Interaction* 29(2): 487–525.

Pardos, Z. A.; and Jiang, W. 2020. Designing for serendipity in a university course recommendation system. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 350–359.

Pardos, Z. A.; and Nam, A. J. H. 2020. A university map of course knowledge. *PLoS ONE* 15(9): e0233207.

Polyzou, A.; Athanasios, N.; and Karypis, G. 2019. Scholars Walk: A Markov Chain Framework for Course Recommendation. In *Proceedings of the 12th International Conference on Educational Data Mining*, 396–401.

Quadrana, M.; Karatzoglou, A.; Hidasi, B.; and Cremonesi, P. 2017. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 130–137.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8): 9.

Ralph, D.; Li, Y.; Wills, G.; and Green, N. 2019. Recommendations from Cold Starts in Big Data. *Proceedings of the 4th International Conference on Internet of Things, Big Data and Security* doi:10.5220/0007798801850194.

Ren, Z.; Ning, X.; Lan, A.; and Rangwala, H. 2019. Grade Prediction Based on Cumulative Knowledge and Co-taken Courses. In *Proceedings of the 12th International Conference on Educational Data Mining*.

Rendle, S.; Freudenthaler, C.; and Schmidt-Thieme, L. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, 811–820.

Sarwar, B.; Karypis, G.; Konstan, J.; and Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, 285–295.

Schuster, M.; and Paliwal, K. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on* 45: 2673 – 2681. doi:10.1109/78.650093.

Sedhain, S.; Sanner, S.; Braziunas, D.; Xie, L.; and Christensen, J. 2014. Social collaborative filtering for cold-start recommendations. In *Proceedings of the 8th ACM Conference on Recommender systems*, 345–348.

Shapiro, D.; and Dundar, A. 2016. Completing College: A National View of Student Attainment Rates. URL https://nscresearchcenter.org/signaturereport12/.

Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 1441–1450.

Sun, M.; Li, F.; Lee, J.; Zhou, K.; Lebanon, G.; and Zha, H. 2013. Learning multiple-question decision trees for cold-start recommendation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 445–454.

Tuan, T. X.; and Phuong, T. M. 2017. 3D convolutional networks for session-based recommendation with content features. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 138–146.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *ArXiv* abs/1706.03762.

Wan, S.; Lan, Y.; Wang, P.; Guo, J.; Xu, J.; and Cheng, X. 2015. Next Basket Recommendation with Neural Networks. *RecSys 2015 Poster Proceedings* .

Wang, S.; Hu, L.; Cao, L.; Huang, X.; Lian, D.; and Liu, W. 2018. Attention-based transactional context embedding for next-item recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Winecoff, A. A.; Brasoveanu, F.; Casavant, B.; Washabaugh, P.; and Graham, M. 2019. Users in the loop: a psychologically-informed approach to similar item retrieval. In *Proceedings of the 13th ACM Conference on Recommender Systems*, 52–59.

Wu, C.-Y.; Alvino, C. V.; Smola, A. J.; and Basilico, J. 2016. Using navigation to improve recommendations in real-time. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 341–348.

Yap, G.-E.; Li, X.-L.; and Philip, S. Y. 2012. Effective next-items recommendation via personalized sequential pattern mining. In *International conference on database systems for advanced applications*, 48–64. Springer.

Yu, F.; Liu, Q.; Wu, S.; Wang, L.; and Tan, T. 2016. A dynamic recurrent model for next basket recommendation. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 729–732.

Zhou, K.; Wang, H.; Zhao, W. X.; Zhu, Y.; Wang, S.; Zhang, F.; Wang, Z.; and Wen, J.-R. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1893–1902.