Statistical and Generative Models with Subtitle Extraction for Next Product Title Generation

Honghee Lee
honghee.lee1101@gmail.com
Sungkyunkwan University
Republic of Korea

Youngjae Chang
youngjaechang0@gmail.com
Sungkyunkwan University
Republic of Korea

Kyuri Choi
gguriskku@gmail.com
Sungkyunkwan University
Republic of Korea

Jongwuk Lee
jongwuklee@skku.edu
Sungkyunkwan University
Republic of Korea

Youngjoong Ko*
yjko@skku.edu
Sungkyunkwan University
Republic of Korea

ABSTRACT

Session-based recommendation aims to predict the next item from the user's actions in the ongoing session. It mainly suffers from the cold start item problem, referring to the difficulty in providing accurate recommendations for items with little or no previous interactions. The KDD Cup 2023 Task 3 (next product title generation) addressed this challenge to improve session-based recommendation. This paper proposes an effective solution for the next product title generation using statistical and generative models. In this process, we optimize a model combination strategy that selects the optimal prediction model for each session based on predefined conditions. The title of the last product serves as a fallback when the session does not meet any conditions. We also devise subtitle extraction techniques to identify a common element among multiple predicted titles. Consequently, our team, *We Bare Bears*, has achieved third place in the KDD Cup Task 3 with a BLEU score of 0.26998, demonstrating the effectiveness of our proposed solution.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

Session-based recommendation, next product title generation, Markov model, generative model, subtitle extraction

ACM Reference Format:
Honghee Lee, Youngjae Chang, Kyuri Choi, Jongwuk Lee, and Youngjoong Ko. 2023. Statistical and Generative Models with Subtitle Extraction for Next Product Title Generation. In *KDD Cup 2023 Workshop: Multilingual Session Recommendation Challenge, August 6, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 4 pages.

*Corresponding author

KDDCup '23, August 6, 2023, Long Beach, CA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.

1 INTRODUCTION

Session-based recommendation [2, 3] predicts the next item from a sequence of previous items consumed by an anonymous user. Its goal is to understand the user's interests in their current session and recommend personalized items. Various real-world applications, such as e-commerce and video streaming platforms, employ session-based recommendation systems. Specifically, E-commerce platforms enhance user shopping experiences through personalized recommendations, directly affecting platform revenue.

To encourage further research in a session-based recommendation, Amazon Search released the Amazon Multilingual Multi-locale Session Dataset (Amazon-M2) [4] and organized the KDD Cup 2023. Task 3 of this competition, next product title generation, presents a unique challenge due to the need to predict cold-start products unseen during the training phase.

In this work, we solve this task by combining three models: a statistical model, a generative model, and a simple fallback method. We first analyze the dataset and choose the final product title of the session as the baseline for fallback, given the prevalence of repeated products within a session. We also observe the strong correlation between consecutive items and leverage it by using Markov Chain as the statistical model. The generative language model estimates the meanings of product titles within a session and generates appropriate titles, enabling robust recommendations, even for items not encountered during the training process. Figure 1 shows our model combination strategy that chooses an appropriate model under specific conditions using item frequency and the confidence of model predictions.

The statistical and generative models primarily employ a greedy search approach, which predicts the title with the highest probability as a complete sequence. In this process, it is observed that the greedy search strategy can incur inaccurate title segments. To overcome this, we enhance the title search process by predicting multiple title candidates and integrating them using a subtitle extraction method. That is, we extract the essential components of the next title by identifying common elements among the title candidates predicted by each model. Given the unique characteristics of the outputs from each model, subtitle extraction is applied separately to each model. By implementing all the proposed methods, our solution has finally achieved a BLEU score of 0.26998.

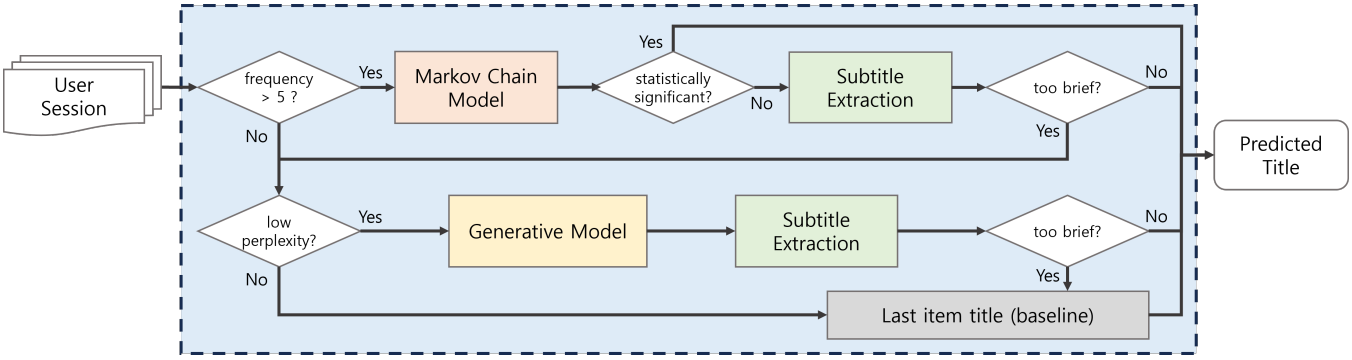

Figure 1: The overview of our architecture. The user's session is processed by our proposed system to predict the next item's title. Note that when we perform subtitle extraction we check if the resulting subtitle is too short, to compensate for brevity penalty.

Table 1: Statistics of the dataset for task 3, Next Product Title Prediction. The acronym ASIN stands for Amazon Standard Identification Number

Language (Locale)	# Train Sessions	# Test Sessions	# Products (ASINs)
German (DE)	1,111,416	10,000	513,811
Japanese (JP)	979,119	10,000	389,888
English (UK)	1,182,181	10,000	494,409
Spanish (ES)	89,047	6,421	41,341
French (FR)	117,561	10,000	43,033
Italian (IT)	126,925	10,000	48,788

2 TASK DESCRIPTION

Table 1 provides the dataset description used in Task 3 (next product title prediction). The dataset is categorized into two parts regarding the volume of resources. The locales with high resources are DE, JP, and UK, while ES, FR, and IT have low resources in both the number of sessions and items. However, the number of test cases is all the same except ES, indicating the importance of performance in languages with low resources. Besides, it is worth highlighting that the number of items in high resources languages is close to 500,000, significantly larger than most recommendation datasets. This causes some limitations in modeling a recommendation approach as it is very inefficient to build an item by an item matrix that is used in many previous recommendation systems [8].

3 DATA ANALYSIS

We analyze the UK locale to find hints for building an optimal model. We count repeated item patterns in the dataset, and it is observed that users tend to click on the same item in the session, which happened in about 10% of the sessions. Moreover, about 5% sessions had cases of the same item clicked on consecutively. Based on this observation, we set the baseline of our work as the title of the last item.

To further understand the dataset's characteristics, we analyze the correlation between session items and the next item. We found that some items have high predictability in choosing the next item. Figure 2 show the results of how prominent the next item is based on the last item in the session. Items that appear less than five times are excluded from the graph to reduce the bias in the trend. The

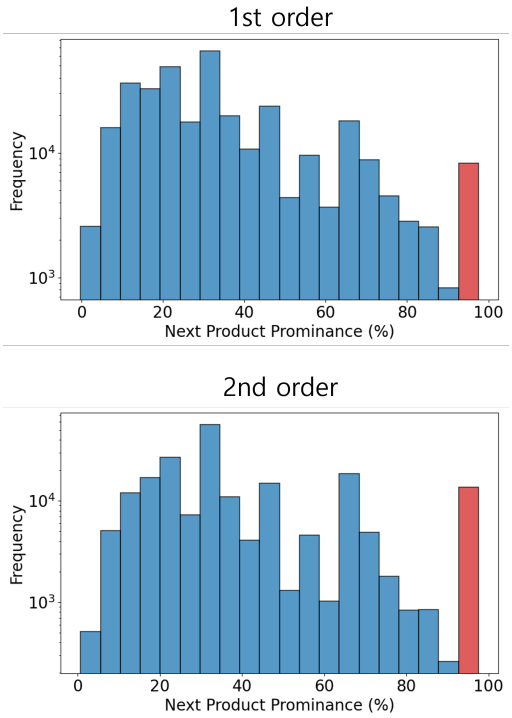

Figure 2: The histogram of the prominence rate for the next item. 1st order notates the conditioning only on the last item while 2nd order conditions on the last two items. The relatively high value of frequency near 100% shows that many items have a highly likely next item.

results reveal that more than 10,000 items out of 200,000 show a 100% tendency of users to click a certain next item. This suggests that simple statistical methods, such as Markov Chain can be an effective solution for a simple case.

4 PROPOSED METHOD

In this section, we present the three models used to predict the next item title. As depicted in Figure 1, each model is applied under certain conditions. If a session does not meet any criteria, the predicted title is defaulted to the last item title, which serves as a strong baseline. Otherwise, we utilize either the Markov Chain model or the generative model.

4.1 Markov Chain Model

Markov Chain is a stochastic model regarding a sequence of events where the probability of each event depends only on the state of the previous event [5]. The task dataset is highly biased by the last one or two items, so conditioning the prediction with Markov Chain will properly address the task's characteristics. Given that $S = \{s_1, s_2, ..., s_n\}$ is a session, where s_i is the i-th item in the session, let s^k_{n+1} be k-th product candidate sorted by descending order for s_{n+1}. $P(s^1_{n+1}|s_n)$ is the transition probability from s_n to s^1_{n+1}, 1st ranked candidate. Then the recommendation R of a first-order Markov Chain is given by:

$$R = \begin{cases} s^1_{n+1} & \text{if } P(s^1_{n+1}|s_n) \geq T \\ SE(s^1_{n+1}, s^2_{n+1}) & \text{otherwise} \end{cases} \quad (1)$$

To apply Markov Chain only when certain, we set a threshold T to check the transition probability. If the probability is smaller than T, the subtitle extraction method is used on the top-2 candidates that will be described in the following section.

4.2 Generative Model

For the generative model, we incorporate a generative pre-trained language model to predict plausible item titles. The primary objective of this approach is to leverage the rich information encapsulated within the pre-trained models and to mitigate the cold start problem. As the foundational model, we use the Text-to-Text Transfer Transformer (T5) [7], a pre-trained transformer model based on an encoder-decoder structure. For the input to the T5 model, we formulate natural language prompts designed to articulate the task at hand, e.g., "Given a list of previously seen item titles, generate the next item title: " Alongside this, we concatenate the most recent three item titles from each session data, divided by '+' signs. Subsequently, the task for the T5 decoder is to generate the next item title based on the prompt input.

To extract the optimal results, we integrate post-processing methods to exploit the top-k model outputs obtained via the beam search algorithm beyond using a single output. A crucial element of this phase is utilizing a subtitle extraction algorithm employed within the Markov Chain model phase (see Figure 1). Moreover, we filter generated outputs based on a particular threshold on the beam search scores, thereby ensuring to include only the outputs generated with high confidence.

4.3 Subtitle Extraction

In addressing the challenge of predicting the desired item title, we focus on predicting subtitles - sequences of words that constitute a portion of the title, with a high likelihood of appearing in the gold title. To extract subtitles, we employ the previously mentioned statistical and generative models to predict candidate titles and then identify their common elements. For subtitle extraction, we utilize the Longest Common Prefix (LCpre) algorithm on generative model outputs and the Longest Common Substring (LCstr) and Longest Common Subsequence (LCsub) algorithms on statistical model outputs.

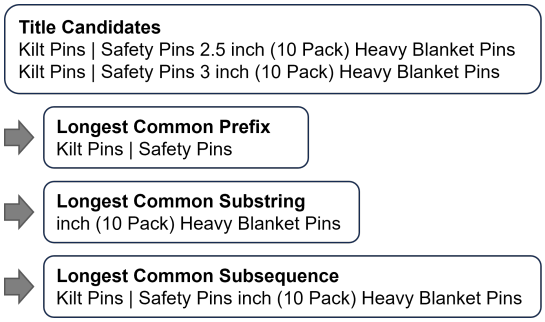

Figure 3: An example of the results from subtitle extraction applied to given title candidates.

The LCpre refers to the longest continuous string of characters shared by the beginnings of two strings. The LCstr denotes the longest string that is a substring of two strings. It represents the longest sequence of characters that appears in each of the strings. The LCseq involves identifying the longest subsequence that is common to two strings. This subsequence comprises a sequence of characters that appears in the same order in each of the strings, but not necessarily consecutively. Figure 3 illustrates examples of applying each common element extraction algorithm.

4.4 Combining the Two Models

We integrate the Markov Chain and the generative model by applying each model to certain conditions. As items that appear less than five times in the train data are usually considered noise, we only apply the Markov Chain model for items with higher frequency. The generative model applies only to the generated titles that fall within the lowest 10% of perplexity scores. If the session does not satisfy any condition, it defaults to the baseline method. In addition, the BLEU score metric has a Brevity Penalty attribute that penalizes predictions with too short lengths. To accommodate this, we discard predictions shorter than certain criteria and return the output using the baseline method. Notably, each of the six locales has a different baseline performance, where low-resource languages have better BLEU performance than high-resource languages. Thus, we set different criteria for each locale to adjust the prediction results.

5 EXPERIMENTS

5.1 Experimental Setup

We describe the following models used in our solution.

- **Base**: It is a simple baseline that uses the title of the last item in the session data.
- **Stat1**: It is a statistical approach implementing the first-order Markov Chain model.
- **Stat2**: It is a variant that incorporates the LCstr on the two candidate items derived from the first-order Markov Chain model.

Table 2: Comparison of baseline and our models on the phase 2 test set.

Id	Method	BLEU	BP
Base	Last item title	0.26553	1
Stat1	Base + 1st-order MC	0.26584	1
Stat2	Base + 1st-order MC + LCstr	0.26878	1
Stat3	Base + 2nd-order MC + LCstr	0.26764	1
Gen1	Base + gen 'UK'	0.26093	1
Gen2	Base + gen 'UK' + LCpre	0.26699	1
Gen3	Base + gen 'UK' + LCsub	0.26616	0.97870
Best	Base + Stat2 + Gen3	**0.26998**	0.99995

- **Stat3**: It applies the LCstr strategy on candidates derived from a Second-Order Markov Chain.
- **Gen1**: FLAN-T5-base model [1], specifically tuned for the 'UK' locale, generates the next item title in the 'UK' session data.[1]
- **Gen2**: It applies LCpre on every beam search output, which was generated through the beam search algorithm.
- **Gen3**: It utilizes the LCsub algorithm to generate plausible outputs.
- **Best**: Our best model combines multiple models.

Evaluation metrics. Unlike traditional recommendation tasks, the prediction task for next product titles employs BLEU score [6], a metric commonly used in natural language generation. The BLEU score assesses the precision by calculating the proportion of correct predicted n-grams to the total predicted n-grams. However, to penalize excessively short predictions, which may achieve high BLEU scores due to their conciseness, the Brevity Penalty (BP) is also utilized for evaluation. The BP adjusts the score downward for overly brief outputs.

5.2 Experimental Results

Table 2 shows the overall performance of the models on the phase 2 test set. Aligned with our data analysis observation of consecutive item repetitions, a simple baseline approach of directly using the last item title yields relatively meaningful results (Base). Furthermore, as can be seen from the results of Stat1, predicting the next item based on significant statistical correlation presents superior performance to the mere employment of the last item. However, when comparing the performances of Stat2 and Stat3, we observe a performance decrease using the second-order Markov Chain technique. This deviation from our data analysis suggests areas for further investigation in future work.

Regarding the generative approaches, when a single output from the fine-tuned T5 model is used directly, denoted as Gen1, it underperforms when compared to the baseline. This result shows the limitations of using a generative model without any post-processing.

A notable result in our experiments is that applying the subtitle extraction algorithm to multiple predicted candidate titles significantly improves the BLEU score, as can be seen by comparing Stat1 with Stat2 and Gen1 with Gen2 and Gen3. Although the performance table indicates the algorithm that most improved performance in each technique, all three subtitle extraction algorithms led to performance enhancements. This outcome demonstrates the efficacy of noise reduction within the predicted titles by subtitle extraction, leading to robust outputs.

Our highest performing method, a combination of models denoted as "Best," achieved a BLEU score of 0.26998, earning us the third position in the KDDCup'23 Challenge Task 3 final leaderboard.

[1]We limited the application of this method solely to the 'UK' since empirical observations revealed that the performance in other locales is suboptimal.

6 CONCLUSION

This paper presents our approach to the Next Product Title Generation task for the KDD Cup 2023. Our solution utilizes statistical and generative models to predict product title candidates and applies subtitle extraction techniques to generate the final predicted title. Our method is based on a thorough analysis of the given dataset. It employs a model combination approach that selects the appropriate model according to the characteristics of the input session. Our solution effectively tackles the cold-start problem and the product repetition within a session. By applying all the proposed methods, we have improved the performance from a baseline BLEU score of 0.26558 to 0.26998. As a result, our solution achieved 3rd place in the KDDCup'23 Challenge Task 3.

ACKNOWLEDGMENTS

We are grateful to both the KDD Cup organizers and Amazon for hosting a great competition. This work was supported in part by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00421, AI Graduate School Support Program (Sungkyunkwan University)).

REFERENCES

[1] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).
[2] Hui Fang, Guibing Guo, Danning Zhang, and Yiheng Shu. 2019. Deep learning-based sequential recommender systems: Concepts, algorithms, and evaluations. In *Web Engineering: 19th International Conference, ICWE 2019, Daejeon, South Korea, June 11–14, 2019, Proceedings 19*. Springer, 574–577.
[3] Dietmar Jannach, Malte Ludewig, and Lukas Lerche. 2017. Session-based item recommendation in e-commerce: on short-term intents, reminders, trends and discounts. *User Modeling and User-Adapted Interaction* 27 (2017), 351–392.
[4] Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, Zhen Li, Monica Xiao Cheng, Rahul Goutam, Haiyang Zhang, Karthik Subbian, Suhang Wang, Yizhou Sun, Jiliang Tang, Bing Yin, and Xianfeng Tang. 2023. Amazon-M2: A Multilingual Multi-locale Shopping Session Dataset for Recommendation and Text Generation. (2023).
[5] James R Norris. 1998. *Markov chains*. Number 2. Cambridge university press.
[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
[8] Harald Steck. 2019. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference*. 3251–3257.