Next Product Title Generation in E-commerce: Rule-based Methods and Autoencoder Model

10th Place Solution for Task 3 in Amazon KDD Cup 2023

Junya Miyamoto
Akasaka Biz Tower
HAKUHODO Inc.
Tokyo, Japan
junya.miyamoto@hakuhodo.co.jp

Shinnosuke Hirano
Akasaka Biz Tower
HAKUHODO Inc.
Tokyo, Japan
shinnosuke.hirano@hakuhodo.co.jp

Soma Makino
Akasaka Biz Tower
HAKUHODO Inc.
Tokyo, Japan
soma.makino@hakuhodo.co.jp

Kazuya Uekado
Akasaka Biz Tower
HAKUHODO Inc.
Tokyo, Japan
kazuya.uekado@hakuhodo.co.jp

Xinyi Lao
Akasaka Biz Tower
HAKUHODO Inc.
Tokyo, Japan
xinyi.lao@hakuhodo.co.jp

ABSTRACT

This paper primarily presents our approach and results for Task 3: Next Product Title Generation in Amazon KDD Cup 2023. Our approach attempted to combine rule-based methods with an auto encoder model. Initially, we set the baseline using the titles of the last item, and then we investigated a rule-based method to generate the next product title based on common words in the titles of products the customer recently interacted with. Subsequently, we experimented with an auto encoder model trained with LSTM, incorporating it with the rule-based approach to generate the next product title. The result demonstrated that the rule-based method achieved the best score. Our approach achieved a 10th place ranking in Task3 in Amazon KDD Cup 2023 competition.

CCS CONCEPTS

• Information systems → Information retrieval → Retrieval tasks and goals → Recommender systems

KEYWORDS

Session-based recommendations, Multilingual recommendation systems, New Cold Start product title generation, KDD CUP

1 Introduction

The utilization of session data and machine learning technologies to predict the next product a customer will purchase based on their preferences is becoming widely adopted. [1-3] However, session-based recommendations that consider multilingual environments and data imbalances have not been thoroughly investigated.
Amazon KDD Cup 2023 proposes to address this issue by advocating the development of multilingual recommendation systems that can enhance personalized recommendations and facilitate an understanding of global preferences and trends. In this competition, a "Multilingual Shopping Session Dataset" is provided, comprising millions of user sessions collected from six different locales where the primary languages of the products are English, German, Japanese, French, Italian, and Spanish.
Amazon KDD Cup 2023 is divided into three tasks. Task 1 involves predicting the product a customer is likely to engage with next, considering the customer's session data and the attributes of each product. The training set for Task 1 is composed of data from locales with relatively large sample sizes English, German, and Japanese. Task 2 is a product prediction task using a smaller sample of data from French, Italian, and Spanish locales. Task 3 offers a new challenge of predicting "cold start" products by generating the title of the product a customer is expected to engage with next, differing from Tasks 1 and 2, which focus on recommendations for existing products.
In this paper, we will discuss our approach to Task 3, in which we were ranked 10th.

KDD'23, Aug, 2023, Long Beach, CA, USA
 978-1-4503-0000-0/18/06...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 METHODOLOGY

In this section, we will introduce the data used in Task 3, the preprocessing methods applied, and the evaluation metrics for the results. Following that, we will explain the main approaches adopted for Task 3.
Task 3 required to predict the title of the next product that a customer will engage with, based on their session data. The test set for Task 3 includes data from all six locales.

The evaluation metric used for Task 3 is bilingual evaluation understudy (BLEU).
BLEU is used to assess the quality of natural language generation by comparing the generated candidate to one or more references. BLEU calculates precision scores for different n-gram lengths and applies a brevity penalty to account for shorter candidates. The final BLEU score is computed as the product of the brevity penalty and the exponential function of the weighted sum of n-gram precision scores. The BLEU-4 score is used in task3 with N=4 and wn=1/N. Higher BLEU scores indicate better generation.

$$BLEU = \min\left(1, \frac{output\ length}{reference\ length}\right)\left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}}$$

Based on the content above, we present a comprehensive analysis of Task 3.
Firstly, we conduct a straightforward preprocessing step on the raw data of Task 3 to ensure data readiness. Next, we establish a baseline approach by using the title of the last item from each customer's session data. Subsequently, we delve into the exploration of a rule-based method. This method is designed to generate the next product title by identifying common words in the titles of products that the customer recently engaged with. The aim is to leverage the correlation observed between the last few products a customer interacted with and the subsequent product they are likely to engage with. Additionally, we venture into the application of a more sophisticated approach, utilizing an autoencoder model trained with LSTM. The goal is to harness the power of deep learning and auto-encoder models in generating product titles that align with the customer's preferences and interactions.
Each of these methods will be elaborated upon extensively in the subsequent sections of this paper.

2.1 Task 3 Data Details and Preprocessing

In this competition, a "Multilingual Shopping Session Dataset" was provided, which contained anonymized session and product information from six different locales: English, German, Japanese, French, Italian, and Spanish. The session information consisted of a list of products that the user engaged with in chronological order, while the product information included details such as the product title, price in local currency, brand, color, and description. The dataset was divided into training and testing sets, with the Task 3 testing set containing products that were not included in the training set.
During the preprocessing phase, we undertook several steps to prepare the data for analysis. Specifically, we removed special characters like brackets, quotation marks, and newline codes from the session information to ensure data cleanliness and consistency. Additionally, we merged the session and product information using the product ID as a key, creating a comprehensive dataset for further analysis. Regarding Task 3, due to the lack of ground truth data as a validation benchmark, we were unable to utilize a separate validation dataset. Instead, we adopted an alternative approach, employing the public score as the evaluation metric. By using the public score, we could assess the performance and effectiveness of our methods.

2.2 Baseline

Before attempting the rule-based method, we established a baseline score as a reference. Through pre-analysis of the training dataset in Task1 and Task2, we found a high correlation between adjacent products that users engaged with. Therefore, we set the baseline score as the score by considering the title of the last product in a customer's session as the generated product title.

2.3 Rule Base Approach

In the rule-based method, we focused on utilizing the titles of the products that the customer engaged with in the last and penultimate actions of their session to generate the title of the recommendation. Our investigation of the training dataset

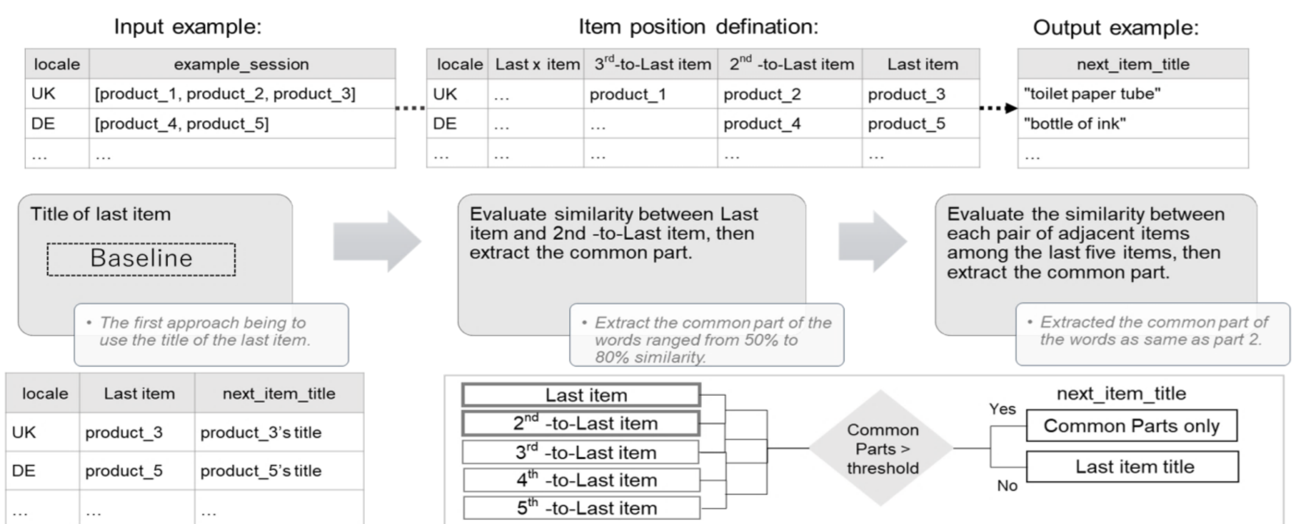

Figure 1. Overview of Rule-based approach

revealed a notable trend that the title of the next engaged item often exhibited similarity to titles of products that had been interacted with repeatedly. As a result, we adopted a rule-based methodology for generating the title of the next recommended item.
The rules we established are as follows: when the titles of products engaged in the last and penultimate actions of the session exhibit similarity, we utilize common elements between these two products to generate the title of the next recommended product. On the contrary, if the engaged products are dissimilar, we directly adopt the title of the last interacted product as the recommendation title for the next item.
The definition of similarity is as follows: considering the characteristics of the BLEU-4 evaluation metric for Task 3, we partition the titles into consecutive units of four words each. Subsequently, we establish a similarity threshold for the adjacent product title units to extract essential content.
Initially, we evaluated similarity and generated titles based on the last item and the second-to-last item in the session. During this stage, we assessed the similarity of product titles by calculating the matching rate of words constituting the titles. We set thresholds at 50%, 60%, 70%, and 80% for generating product titles and calculating their corresponding scores.
Subsequently, we further investigated the generation of titles by considering the last product in the session and the five products preceding it. We generated titles based on this set of products and calculated their scores if there were any similar items among them. The similarity threshold used in this context was determined based on the value that resulted in the highest score, as established in the approach utilizing the last and second-to-last items in the session.
Figure 1 shows an overview of this approach.

2.4 Generation Approach Using Auto-Encoder

The rule-based approach we proposed in Section 2.3 is an approach that generalizes and recommends the name of the product that the user engaged with at the end of the session. In the rule-based method, we thought there was a high possibility of falling into a local solution because it generalizes using only the product information in the same session. Therefore, we attempted to remove noise and generalize by constructing an Auto-Encoder [4] model that reproduces the last product title of the session.
Our Auto-Encoder model is composed of two main layers in both the encoder and decoder: an embedding layer and an LSTM (Long Short-Term Memory) [5] layer. The encoder takes the preprocessed text sequences as input. These sequences are first passed through an embedding layer, which transforms the integer-encoded vocabulary into dense vector representations. The output of the embedding layer is then fed into the LSTM layer, which processes the embedded sequences and outputs a set of 'encoder states'. These encoder states represent the LSTM's learned understanding of the input sequences.
Like the encoder, the decoder is also composed of an embedding layer and an LSTM layer. The LSTM layer of the decoder processes the sequences based on the initial states and generates the output sequences. The overview is shown in Figure 2.
We used the product title generated by the Auto-Encoder model, which was input with the product that the customer engaged with at the end of the session. However, as in Section 2.3, we evaluated the similarity based on the last product and the second to last product of the session, and implemented this process only for sessions that have a certain similarity between the two.

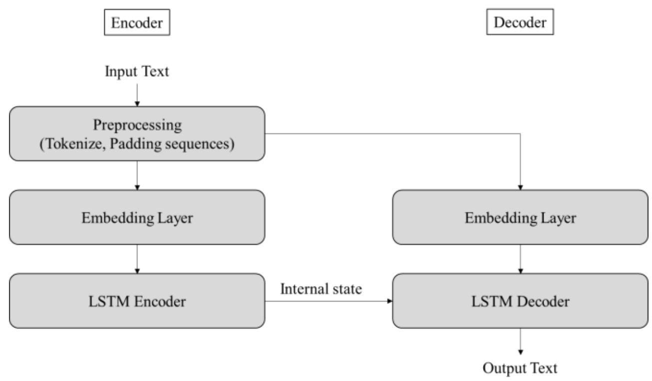

Figure 2. Overview of Auto-Encoder approach

3 Result

We show our results in Table 1. The score of the baseline approach, which considered the name of the product that the customer last engaged with as the generated product title, was 0.26553 on the public leaderboard. When generating product titles based on the common words in the titles of the product the customer last engaged with and the second to last, some scores were improved. In particular, the best score of 0.26787 was recorded when recommending with product titles generated when the match rate of words in the two products was more than 70%. However, the approach of generating titles based on a product similar to the last product of the session and any of the previous five products did not contribute to accuracy improvement.
Also, when implementing an approach using Auto-Encoder, the accuracy improved against the baseline, but it did not reach the best score (BLEU=0.26696). One possible reason for this is that the generation of symbols was not successful. Although we were not able to fully implement it in this competition, we assume that tuning the decoder hyperparameters could potentially improve this.

Table 1. Some results of our experiments

Approaches	BLEU Score	Brevity Penalty
Baseline: Last Item's titles Only	0.26553	1.00000
Rule Base1: 50% common	0.26019	0.96430
Rule Base2: 60% common	0.26497	0.98584
Rule Base3: 70% common	0.26787	1.00000
Rule Base4: 80% common	0.26695	1.00000
Rule Base5: Last5Session common	0.24915	0.99195
Encoder-decoder generation based on Rule Base3	0.26696	0.99676

4 Conclusion

In this paper, we detailed our solution to Task 3: Next Product Title Generation of the Amazon KDD Cup 2023. We generated new product titles from the common words in the titles of the product the customer last engaged with and the one before last, and achieved a ranking of 10th place.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to the organizing teams from Amazon and AIcrowd for their tremendous efforts throughout the entire competition. Hosting this remarkable event has provided us with invaluable experiences. We are truly grateful for the passion and dedication put forth towards achieving success. We would also like to extend our heartfelt thanks to everyone involved in organizing and sponsoring the KDD Cup 2023.

REFERENCES

[1] Adomavicius, Gediminas, and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering, 17(6), 734-749.
DOI: https://doi.org/10.1109/TKDE.2005.99

[2] Lops, Pasquale, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. Recommender systems handbook, 73-105.
DOI: https://doi.org/10.1007/978-0-387-85820-3_3

[3] Hidasi, Balázs, et al. 2015. Session-based recommendations with recurrent neural networks. arXiv preprint arXiv:1511.06939.
DOI: https://doi.org/10.48550/arXiv.1511.06939

[4] Hinton GE, Salakhutdinov RR. 2006. Reducing the Dimensionality of Data with Neural Networks. Science, 313(5786), 504-507.
DOI: https://doi.org/10.1126/science.1127647

[5] Sepp Hochreiter, Jurgen Schmidhuber. 1997. Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.
DOI: https://doi.org/10.1162/neco.1997.9.8.1735