Word Filtering Approach for Next Product Title Generation

Yusuke Fukushima*
NTT DOCOMO, INC.
Tokyo, Japan
yuusuke.fukushima.fw@nttdocomo.com

Keiichi Ochiai*
NTT DOCOMO, INC.
Tokyo, Japan
ochiaike@nttdocomo.com

Yoshitaka Inoue
DOCOMO Innovations, Inc.
Sunnyvale, CA , USA
yoshitaka.inoue@docomoinnovations.com

Masato Hashimoto
NTT DOCOMO, INC.
Tokyo, Japan
masato.hashimoto.px@nttdocomo.com

Sho Maeoki
NTT DOCOMO, INC.
Tokyo, Japan
syou.maeoki.rz@nttdocomo.com

ABSTRACT

The rise of e-commerce has prompted e-commerce platform providers to make efforts toward improving the shopping experience of their users. One approach involves the optimization of product recommendations tailored to user preferences and the presentation of product titles based on historical session data, which encompass past instances of browsing various products. In considering these, the organizers of KDD Cup 2023 launched a competition focused on session-based recommendations, accompanied by a comprehensive dataset known as the Amazon Multilingual Multi-locale Session Dataset (Amazon-M2). In this paper, we, NTT-DOCOMO-LABS-RED, present our solution that achieved a 6th place ranking on the public leaderboard for Task 3 (Next Product Title Generation) in the competition. The proposed solution centers around the key notion that words appearing infrequently in product titles are likely extraneous and introduce noise to the prediction. By identifying and removing such words, we aim to enhance the overall quality of the solution. Furthermore, we introduce a second solution that attained a ranking equivalent to 10th place on the public leaderboard. The core idea behind the second solution involves incorporating a prediction for the BLEU score by submitting the last title to the first solution. This prediction is then utilized to determine whether to retain the best solution as it is or to apply further modifications to the title.

CCS CONCEPTS

• **Applied computing** → **Electronic commerce**.

KEYWORDS

KDD Cup, Multilingual Session Recommendation, Text Generation

ACM Reference Format:
Yusuke Fukushima, Keiichi Ochiai, Yoshitaka Inoue, Masato Hashimoto, and Sho Maeoki. 2023. Word Filtering Approach for Next Product Title Generation. In *KDDCUP '23: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 6–10, 2022, Long Beach, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Equal contribution

KDDCUP '23, August 6–10, 2022, Long Beach, CA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the proliferation of e-commerce, e-commerce platform providers are trying to improve the shopping experience of their users. For example, improvements to the user experience are expected if a platform can recommend products that match the user interests or present product titles based on a series of past product browsing histories (referred to as session data). Although session-based recommendations have been studied for some domains such as e-commerce products and videos [4], limitations still remain such as the language of the data is limited and providing a relevant product title, which is important information as a factor leading to a click. To this end, the organizer of KDD Cup 2023 provided the following three tasks and a dataset referred to as the *Amazon Multilingual Multi-locale Session Dataset (Amazon-M2).*

- Task 1: Next Product Recommendation
- Task 2: Next Product Recommendation for Underrepresented Languages/Locales
- Task 3: Next Product Title Generation

The provided data are a collection of anonymized customer sessions containing products from six different locales [5].

In this paper, we introduce a solution that placed 6th on the public leaderboard for Task 3 in the competition. The key concept for the solution is that if a word appears only a few times in all product names, it is probably an unnecessary word that can be considered noise. In addition, we introduce a second solution that is equivalent to 10th on the public leaderboard. The key concept for the second solution is to add a prediction for the Bilingual Evaluation Understudy (BLEU) [7] score if the last title is submitted to the first (best) solution, and use this prediction to determine whether to take the best solution as it is or to modify further the title.

2 RELATED WORK

Session-based Recommendations. Session-based recommendation is a research field receiving attention recently [11]. According to survey paper [11], a unique characteristic of session-based recommendations is to capture dynamic user preferences in the short-term compared to existing recommendations such as collaborative filtering and content-based recommendation, which usually model

long-term yet static user preferences. In session-based recommendations, Recurrent Neural Networks (RNNs) are often used as the architecture. In addition to RNN-based architectures, Transformer-based models are have been proposed such as in [9] and [2] based on the success of Transformers [10] in natural language processing (NLP) tasks, *e.g.*, next word prediction, language translation, and summarization. However, no existing study has considered the task of next product title prediction.

Natural Language Processing. Because the task of next product title prediction is related to text generation tasks, we briefly review the related studies. A straightforward strategy in recent years is to fine-tune a generative pre-trained language model such as BERT [3] and Generative Pre-trained Transformer (GPT) [8] using task-specific training data [6]. This approach is referred to as "pre-train then fine-tune" [6]. Although pre-trained language models have been used for a variety of tasks, including question answering and task-specific prediction, none have been used to predict product titles.

3 APPROACH

Task Description. Task 3 which we introduce our solution is specifically defined as: *Given a session data and the attributes of each product, the goal of this task is to predict the title of the next product that a customer will engage with*[1].

3.1 First Solution: Word Filtering Approach

In Task 3, we can achieve a good score by using the last product title in each session as the inference result. We improved the score by removing infrequently appearing words based on this method. If there is no change in the number of n-grams matched by the correct answer and the inference result, the BLEU score becomes better with fewer tokens in the inference result. Despite the large number of product titles included in the current product data, there are a number of words that appear in only one product title. If we make the assumption that past sessions do not contain the correct product, then the last product in the session, *i.e.*, the product we use as the inference result, is different from the actual correct product. If this assumption is correct, then the BLEU score will always improve by removing these words used in only one product title from the inference results, since a word that appears in only one product will never be included in both the inference results and the correct product.

The next point that needs to be considered is that if the inference result is shorter than the correct answer, the BLEU score will be penalized. When deleting words from the inference result, it is necessary to investigate which language products can be prioritized to delete words from the inference result to improve the score further without penalty. To investigate this, we set the following two parameters for each language. In total, we used 12 parameters.

- Remove words that appear only in n or fewer product titles.
- When deleting a word that appears only in the above n products, delete it with a probability of p.

Note that words that appear only in $n-1$ or fewer products are deleted with 100 percent probability regardless of the value of p.

[1] https://kddcup23.github.io/

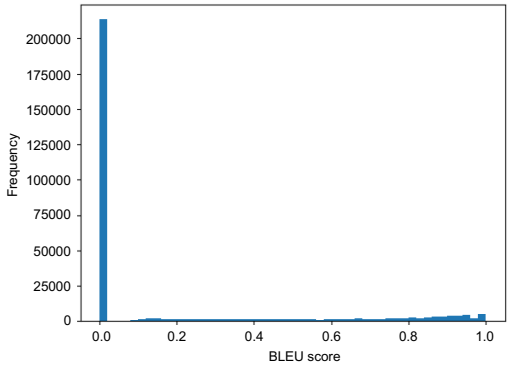

Figure 1: Distribution of BLEU scores if the last product title in a session is used as a prediction

This is because deleting words that appear in only $n-1$ products is clearly less likely to affect the number of n-gram matches between the inference result and the correct answer than deleting words that appear in only n products.

3.2 Second Solution: Combining Word Filtering Using BLEU Score Prediction Model

First, because we would like to discern a tendency in the case that the last product title in a session is a prediction, we calculate BLEU scores for training data with the last title in a session as a prediction. Figure 1 shows the distribution of BLEU scores if the last product title in a session is used as a prediction. From this figure, we hypothesize that if we can classify whether or not the BLEU score is 0, the prediction performance can be improved for the method that uses the last product title in a session as a prediction. Thus, we build a supervised machine learning (ML) model to predict whether or not the BLEU score is 0 from session information. The label is 1 if the BLEU score in the case that the last product title in a session as a prediction exceeds a certain threshold, and 0 otherwise. We use XGBoost [1] as a supervised ML model and the following session information is used as features:

- Session length
- Length of product titles in a session
- Difference in product titles between consecutive products in a session
- Average and standard deviation of product title length in a session

We design the second solution as described in Algorithm 1. First, the BLEU score is predicted from session information (line 1). The prediction probability threshold for binary classification is set higher than 0.5 to emphasize precision because we would like to capture more cases where the model predicts that the BLEU score is 0 and the actual BLEU score is also 0. Then, if the predicted BLEU score is 0 and the token length of the last product title in a session as a prediction is less than or equal to TH_{len}, then the second last product title is set as the (temporary) predicted title (line 3); otherwise, the last product title is set as the (temporary) predicted title (line 5). Finally, we apply the word filtering approach

Algorithm 1 Procedure of the second solution for a session

Input: Term Frequency Dict *Dict*, BLEU score prediction model *Model*, Session information *s*
Output: Next Product Title *pred_title*

1: $pred\_label \leftarrow Model(s)$
2: **if** $pred\_label = 0$ **and** $token\_len(s.last\_title) <= TH_{len}$ **then**
3: $pred\_title \leftarrow s.second\_last\_title$
4: **else**
5: $pred\_title \leftarrow s.last\_title$
6: **end if**
{// The following process is equal to the first solution}
7: **for** word in pred_title **do**
8: **if** word appeared in *Dict* **then**
9: $r \leftarrow random()$
10: **if** $r < TH_{random}$ **then**
11: remove the word from *pred_title*
12: **end if**
13: **end if**
14: **end for**
15: **return** *pred_title*

Table 1: Statistics for provided dataset

Language (Locale)	# Sessions	# Products (ASINs)
German (DE)	1111416	513811
Japanese (JP)	979119	389888
English (UK)	1182181	494409
Spanish (ES)	89047	41341
French (FR)	117561	43033
Italian (IT)	126925	48788

described in Section 3.1. Term TH_{random} in line 10 is the same as TH_{lang} described in Section 4.2.

4 EVALUATION

4.1 Evaluation setting

Dataset. The KDD Cup organizer provided a dataset, referred to as Amazon-M2, for this competition [5]. The dataset includes user sessions containing products from six different locales (English, German, Japanese, French, Italian, and Spanish). User sessions comprise a sequential collection of products with which a user has interacted, organized in chronological order. On the other hand, product attributes include various details such as product title, price in local currency, brand, color, and description. Each product can be identified by a unique Amazon Standard Identification Number (ASIN). The statistics of the provided dataset are given in Table 1.
Metric. The evaluation metrics for this task is BLEU score [7]. BLEU is a metric of the similarity of sentences that is often used for performance evaluation in machine translation.

4.2 Evaluation of First Solution

Two parameters for each language are described in Section 3.1. For simplification of the notation, parameter p is expressed by adding the value of its probability to $n-1$. For example, if $n = 2$ and $p = 0.5$, it is expressed as n=1.5. As another example, for $n = 1$ and $p = 1$, it is represented as $n = 1$. From the definition of the two parameters, this manner of expression is a natural representation given that the first solution results are equal when $n = 1$ and $p = 1$, and when $n = 2$ and $p = 0$. In this section, we define $TH_{lang} = (n-1) * p$ to represent the above calculation. The parameters for each language are expressed by adding the abbreviation of the language to TH_{lang} such as TH_{JP} for a Japanese product.

Table 2: Results of first solution in phase 2 leaderboard

TH_{DE}	TH_{ES}	TH_{FR}	TH_{IT}	TH_{JP}	TH_{UK}	submit score
0	0	0	0	0	0	0.26553
1	1	1	1	1	1	0.26135
1	0	0	0	0	0	0.26617
0	1	0	0	0	0	0.26643
0	1	1	0.1	0	1	**0.26906**
0	1	1	0.15	0	1	0.26895

Examples of the submission results in Phase 2 are given in Table 2. Because of the switch from Phase 1 to Phase 2 during the investigation of the impact of each language, we were not able to conduct a consistent investigation. However, for Spanish, French, and English there was a significant improvement in score using this method. On the other hand, the scores for German and Japanese did not improve even when large numbers of words were removed, and these languages were relatively easy to penalize. We believe that this is a language-specific problem of having many proper nouns and coined words. Italian exhibited intermediate characteristics.

Finally, the best conditions were found to be a combination of the following parameters and the best score is 0.26906.

- The values of $TH_{ES}, TH_{FR}, TH_{UK} = 1.0$
- The value of $TH_{IT} = 0.1$
- The values of $TH_{JP}, TH_{DE} = 0.0$

4.3 Evaluation of Second Solution

Here, we describe three evaluations for the second solution. More specifically, evaluations of the (1) BLEU score prediction model, (2) offline evaluation, and (3) online evaluation for next product title generation.
(1) Evaluation for BLEU score prediction model. First, we randomly sampled 300,000 sessions from the published dataset. Then, we split the sampled dataset into training and validation groups. The training data were randomly sampled from 70% of the sampled dataset and the remaining 30% was used as the validation data. We set the label threshold to 0.1. The label ratios were imbalanced: we had 28.4% for label 1 (BLEU score >= 0.1) and 71.6% for label 0 (BLEU score < 0.1) in total. We empirically tuned the threshold for binary classification, and finally set the threshold to 0.8. Under these settings, precision was 0.889 and recall was 0.257 for BLEU score prediction.
(2) Offline evaluation. For validation data, we compared BLEU scores among three methods: (1) the last product title in a session as a prediction (Baseline), (2) the first proposed solution, and (3) the second proposed solution. The evaluation results are given in Table 3. The results suggest that the second solution using BLEU score prediction may be effective.

Table 3: Results of the offline evaluation

Method	BLEU score (offline)	BLEU score (online)
Baseline	0.18083	0.26553
First solution	0.18107	0.26906
Second solution	0.18125	0.26813

(3) Online evaluation We compared BLEU scores among the three methods described above in the online setting by submitting them to the competition platform. The evaluation results are also given in Table 3. The scores are BLEU scores on the public leaderboard. Unfortunately, although the second solution cannot outperform the first solution, the score (0.26813) is equivalent to 10th on the public leaderboard.

5 CONCLUSIONS

In this paper, we introduced solutions for Task 3 in the competition. The first solution adopted a word filtering approach based on the key concept that if a word appears only a few times in all product names, it is probably an unnecessary word that can be considered noise. The first solution placed 6th on the public leaderboard. The second solution added a BLEU score prediction model to try to improve the prediction performance. The performance of the second solution was equivalent to 10th on the public leaderboard.

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[2] Gabriel de Souza Pereira Moreira, Sara Rabhi, Jeong Min Lee, Ronay Ak, and Even Oldridge. 2021. Transformers4rec: Bridging the gap between nlp and sequential/session-based recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 143–153.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[5] 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).

[6] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *Comput. Surveys* (2021).

[7] 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.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[9] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 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.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[11] Shoujin Wang, Longbing Cao, Yan Wang, Quan Z Sheng, Mehmet A Orgun, and Defu Lian. 2021. A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)* 54, 7 (2021), 1–38.