Prediction of next product title by simple deletion of the last word

Meisaku Suzuki*
NTT DOCOMO, INC.
Tokyo, Japan
meisaku.suzuki.fw@nttdocomo.com

Ryo Koyama*
NTT DOCOMO, INC.
Tokyo, Japan
ryou.koyama.aw@nttdocomo.com

Shin Ishiguro
NTT DOCOMO, INC.
Tokyo, Japan
shin.ishiguro.tb@nttdocomo.com

Shunya Imada
NTT DOCOMO, INC.
Tokyo, Japan
syunya.imada.dn@nttdocomo.com

Yuya Kunimoto
NTT DOCOMO, INC.
Tokyo, Japan
yuuya.kunimoto.ah@nttdocomo.com

ABSTRACT

This paper presents the 9th rank method in KDD Cup 2023 task 3. In the Product Title Prediction Task of KDD Cup 2023 task 3, we improved the BLEU score by removing words in the second half of the product title, as opposed to the baseline method that uses the last product title as the predicted title, which may contain useless words. Since the second term of the BLEU formula penalizes product titles with long word counts, we improved BLEU by optimizing the second term of BLEU by removing words in the second half of the product title. The result was a 0.00268 improvement in BLEU over the baseline method, resulting in a final BLEU of 0.26821, which placed 9th on the Leader Board.

CCS CONCEPTS

• Applied computing → Electronic commerce

KEYWORDS

KDD Cup, Multilingual Session Recommendation

ACM Reference format:

Meisaku Suzuki, Ryo Koyama, Shin Ishiguro, Shunya Imada, Yuya Kunimoto. 2023. Prediction of next product title by simple deletion of the last word. In KDDCUP '23: ACM SIGKDD Conference on Knowledge Discov- ery 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

1 Intrudution

In KDD Cup task3, we worked on a task to predict the next product title in the Amazon dataset. Predicting product titles, as opposed to predicting by product ID, has the potential to address the cold start problem of products. For example, if there is a new product and the new product is similar to the product title predicted for each user, then recommending the new product to the user may result in that product being viewed. Baseline methods that output the last title as is are powerful, and in our experiments, approaches such as product title prediction using product ID prediction and product text generation were used, but they conversely lowered BLEU. Here, we found that EDA contained many useless words in the second half of the product title. Therefore, we propose a method that uses two rules to remove useless words, which improves the BLEU by 0.00268 BLEU compared to the baseline method. This method has the advantage of being a very simple rule-based method, which is highly interpretable, requires fewer compute resources, and takes less time for training and inference. On the other hand, it may be vulnerable to dataset drift. In this method, we proposed Rule 1, which significantly reduces the number of words, because the longer the title is, the worse the score becomes due to the nature of the BLEU score. We also proposed Rule 2, which reduces the word count by a small number of words even for medium-length titles because the end of the title is a keyword for SEO. As a result, we obtained BLEU 0.26821, which is equivalent to KDDCUP task3 9th rank..

2 DATASET AND EXPLORATION

2.1 Dataset

Dataset. The KDD Cup organizer provides a dataset, called Amazon-M2, for this competition[1].

2.2 Exploratory analysis

Next, a histogram of the number of words in product titles is shown in Figure 1. From Figure 1. it can be seen that only a small fraction of the total data has product titles larger than 30-40 words, and the histogram has a skewed distribution. This may indicate that the most of the product titles may not be correct in case if the predicted title has longer than 30-40 words.

First, Table1 is a sample of product title data extracted from the English (UK) dataset of KDD Cup 2023. Words that essentially describe the product, such as the product name, are placed in the first half of the title. On the other hand, qualifying words (e.g. "Shockproof", color ("Navy", "Pink", "Purple"), size "60 x 90 cm",

∗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

Product title
ROCKBROS Bike Mudguards Set, Adjustable 2 Part Mountain Bike Mud Guard Full Cover Universal Front and Rear Bicycle Fender for 22-26.4mm Front Fork
DEXI Dirt Trapper Door Mat,Non-slip Barrier Mats for Indoor and Outdoor,Super Absorbent Entrance Rug Machine Washable Soft Floor Mat Carpet(Blue-Black,60 x 90 cm)
Flexible and Soft Case Cover for 10.9" iPad Air 5 2022 and iPad Air 4 2020. Protects Your Tablet and has a Built In Kickstand and Holder for a Pencil. Colour is Navy
OtterBox Symmetry+ Case for iPhone 13 with MagSafe, Shockproof, Drop proof, Protective Thin Case, 3x Tested to Military Standard, Antimicrobial Protection, Pink
PULEN for Samsung Galaxy S22 Case, Slim Liquid Silicone Shockproof Protective Case - Purple

Table1. Sample product titles for English (UK) dataset

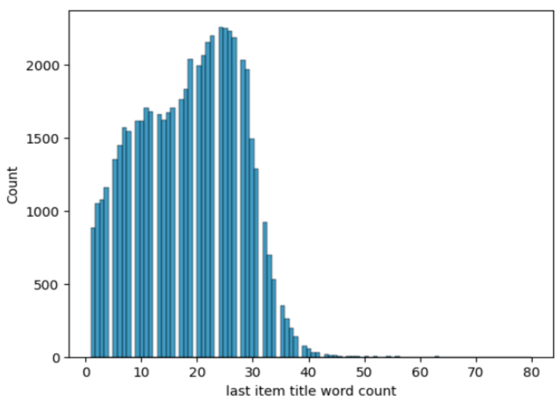

Figure1. Histogram of word counts for product titles

etc.) tend to be placed in the second half. The reason titles are created in this order is thought to be because they are placed with Amazon's SEO strategy in mind. In other words, he first half of the product title that the user sees, while the second half of the product title is intended to get as many hits as possible in a product search.

Finally, the formula for the evaluation metric BLEU[2] is

$$BLEU = BP_{BLEU} + \exp\left(\sum_{n=1}^{N}(w_n \log p_n)\right)$$

where the second term of BLEU, p_n, is as follows

$$p_n = \frac{\sum i \text{ Number of n} - \text{grams matched by translation i and reference translation}}{\sum i \quad \text{Total number of n} - \text{grams in translation i}}$$

From the second term p_n of BLEU above, it is assumed that the BLEU score worsens when the predicted product title is penalized with a long word count.

2.3 Solution

Based on the EDA results described so far, in addition to the baseline approach of adopting the last seen product title as a prediction, we adopted an approach of removing the second half of the product title in order to optimize the second term of the BLEU. Finally, we applied the following two rules and three hyperparameters, *x*, *y*, and *z*, to all multilingual data sets.

- Rule 1 : If the number of words in a product title is less than *y*, delete the *x*-th and subsequent words of the product title.
- Rule 2 : If the number of words in the product title is greater than or equal to *y*, remove *z* numbers of words from the end of the title

Rule 1 is designed to be adapted for a larger number of targets and is used to delete words in the product text. For outlier product titles with too many words, the words after number *x* are deleted. Product title words that were outlier-like with too many words were removed after the number *x* using a threshold. On the other hand, Rule 1 may contain useless words in the product title even within the word count of Rule 1, so a rule is needed to remove those useless words. For this reason, Rule 2 aims is a complementary rule for cases that cannot be addressed by to delete local words that cannot be supplemented by Rule 1. Even if the number of words is small, we observed many titles with SEO words at the end. There for we removed z words after the number y because it is considered to have SEO words at the end..

3 Evaluation

In the evaluation, we will compare the changes in BLEU scores of generated titles to which we applied baseline method, Rule 1 and Rule 2 based on the solution presented in Section 2.3. As a baseline method, we use the method of outputting the title of the last product in the list of the product title search query. First, as an evaluation of Rule 1, we removed words from the title according to Rule 1. The BLEU scores for each number of words removed are shown in Figure 2. The result show that applying word deletion based on this simple rule improves BLEU over the baseline method. Next, to evaluate the word deletion method Rule 2. we conducted an experiment to tune BLEU scores by fixing the parameter *z* is 1 and changing *x, y*. Figure 3 shows the evaluation results. In our experiment, we observed that the BLEU score takes the larger value when *x* was set to 30 or 31 than the others. In the final solution, we set *x*:30 for rule 1, *y*:26 and *z*:1 for rule 2. (This parameter is what our team ranked 9th in the competition)

Finally, Table 2 shows the BLEU scores of our proposed method compared to the baseline method. In Rule1 BLEU score, *x* is setted 28 which is best BLEU score. And in Rule2 BLEU score, *x, y* and *z* are setted 30, 26 and 1 which is best BLEU score.

4 Discussion

The parameters we set in the final solution are the result of repeated registration on the public leaderboard during the final period of the competition period. For this reason, the parameter search could not be carried out sufficiently exhaustively. Therefore,

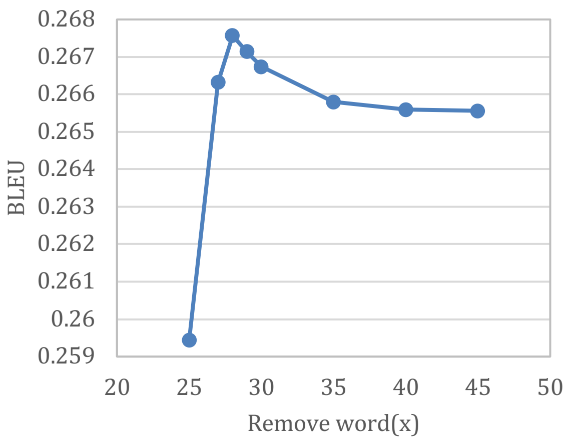

Figure2. BLEU per x number of words deleted of Rule1

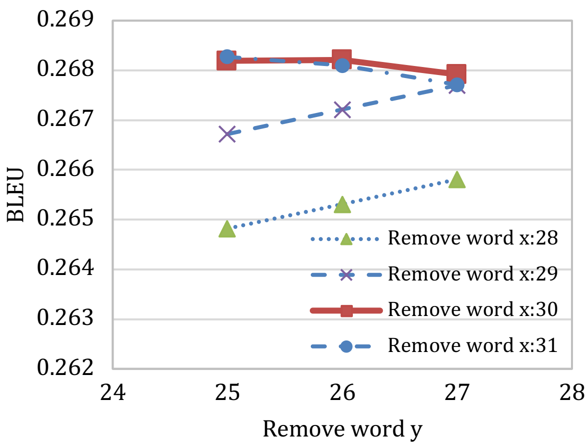

Figure3. BLEU per word deletion number x,y of Rule2

Method	BLEU
Baseline	0.26553
Rule1	0.26757
Rule1+2	0.26821

Table2. Result of BLEU score

there is room for improvement with further, exhaustive hyperparameter searches. If the number of x,y is very different, it may be possible to improve BLEU by setting z to a value other than 1 (2, 3, 4, etc.), but in our experiments, we set x,y to close values, so z was set to 1 in our experiments. However, it may be possible to improve it further by exhaustively changing x, y, and z parameter.

As an alternative approach is by using machine learning to learn and infer the number of word deletions x, y, and z. We think this approach will also improve the BLEU value and make it more robust against data drift.

5 Conclusion

We worked on KDD Cup 2023 task 3, predicting the next product title against Amazon's multilingual shopping dataset. After conducting EDA, we hypothesized the following: Words at the end of item titles are for SEO purposes and are believed to have a negative impact on BLEU scores in the next title prediction. Based on this hypothesis, we designed a simple rule technique for removing words at the end of titles generated from previous search query titles. We experimented with removing words based on two rules for optimizing BLEU. As a result, we obtained a BLEU score of 0.26821, which is 9th place on the leaderboard..

REFERENCES

[1] Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu 90 Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, Zhen Li, Monica Xiao Cheng, Rahul 91 Goutam, Haiyang Zhang, Karthik Subbian, Suhang Wang, Yizhou Sun, Jiliang 92 Tang, Bing Yin, and Xianfeng Tang. 2023. Amazon-M2: A Multilingual Multi-locale 93 Shopping Session Dataset for Recommendation and Text Generation. (2023).

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu IBM T. J. Watson Research Center Yorktown Heights, NY 10598, USA. BLEU: a Method forAutomatic Evaluation of Machine Translation