KDD Cup 2023
Task 3 6th place

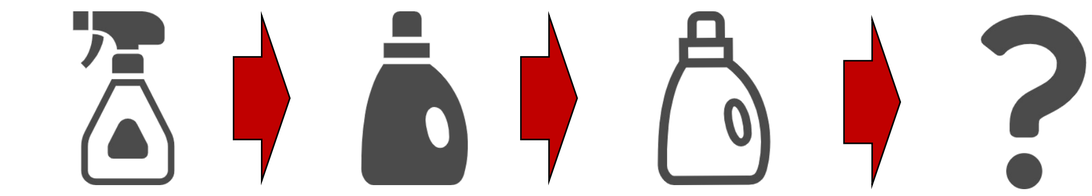
Word Filtering Approach for Next Product Title Generation

Yusuke Fukushima[1,*], Keiichi Ochiai[1,*], Yoshitaka Inoue[2], Masato Hashimoto[1], Sho Maeoki[1]
[1] NTT DOCOMO, INC. [2] DOCOMO Innovations, Inc.
*Equal contribution

Task Description

Task 3: Next Product Title Generation

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



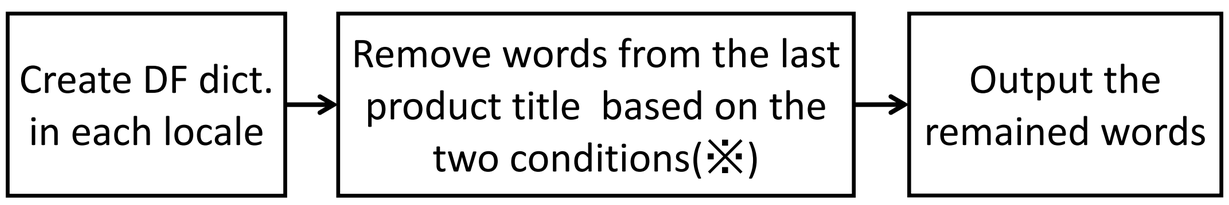
First Solution: Word Filtering Approach

The key ideas

1. If a word appears only a few times in all product names (i.e., document frequency (DF) is low) , it is probably an unnecessary word that can be considered noise
2. The last product title in each session achieves a good score

→ We improved the score by removing infrequently appearing words from the last product title

Overview



※Conditions

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

Evaluation of First Solution

- Because n and p can be expressed as integer and fractional parts of a single parameter, We define $TH_{lang} = (n-1) + p$ to merge the DF
- We submitted several patterns of thresholds as shown in the following table
- This method is effective for Spanish, French, and English, but not for German and Japanese

TH_{DE}	TH_{ES}	TH_{FR}	TH_{IT}	TH_{JP}	TH_{UK}	LB 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	0.26617
0	1	1	0.1	0	1	**0.26906**
0	1	1	0.15	0	1	0.26895

Who we are

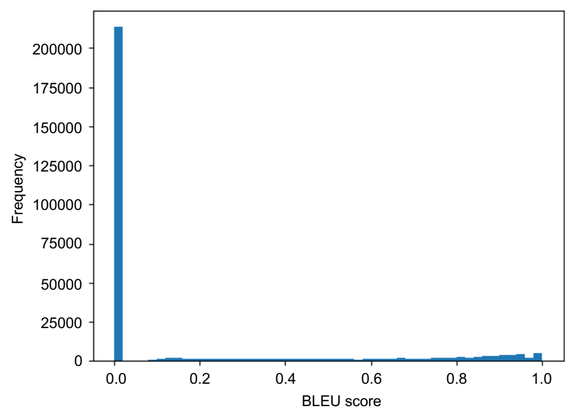
- NTT DOCOMO is a largest mobile operator in Japan
- Our KDD Cup history
  - ✓ 2016 Finalist
  - ✓ 2019 1st Place (Regular ML Track)
  - ✓ 2020 3rd (RL Track), 4th (ML Track2) and 7th (ML Track1) place
  - ✓ 2021 8th place (OGB Large-Scale Challenge)
  - ✓ 2022 9th place (ESCI Challenge Task 3)
  - ✓ 2023 6th place (Task 3)

Second Solution: Combining Word Filtering Using BLEU Score Prediction Model

The key idea

If we can classify whether or not the BLEU score is 0, the prediction performance can be improved based on the visualization of distribution of BLEU scores

→ We combine BLEU score prediction and first solution



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

BLEU score prediction

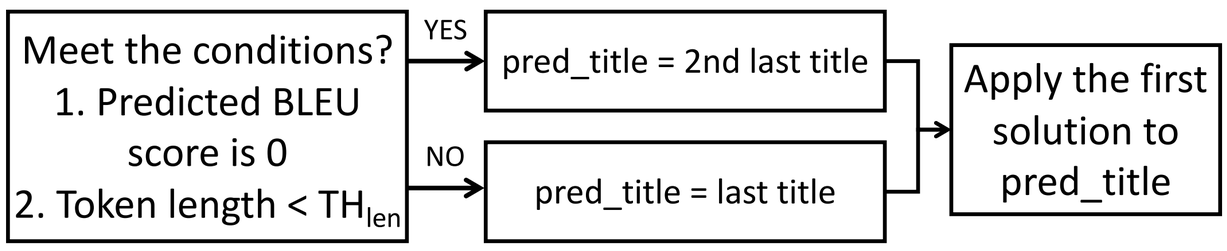
Model: XGBoost

Label: 1 (BLEU score>TH_{BLEU})
0 (Otherwise)

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

Overview



Evaluation of Second Solution

(1) Evaluation for BLEU score prediction model

Data: Randomly sampled 300,000 sessions
(Train 70%, Validation 30%)

Label: 28.4% for label 1 (BLEU score >= 0.1) and 71.6% for label 0

Precision 0.889, Recall 0.257

(2) Offline and online evaluation

We compare the following three methods, (1)Baseline: the last product title in a session as a prediction, (2) First solution, and (3) Second solution. The score (0.26813) is equivalent to 10th on the public leaderboard

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