Convolutional Neural Network and Bidirectional LSTM Based Taxonomy Classification Using External Dataset at SIGIR eCom Data Challenge

Shogo D. Suzuki
Yahoo Japan Corporation
Tokyo, Japan
shogosu@yahoo-corp.jp

Yohei Iseki
Yahoo Japan Corporation
Tokyo, Japan
yiseki@yahoo-corp.jp

Hiroaki Shiino
Yahoo Japan Corporation
Tokyo, Japan
hshiino@yahoo-corp.jp

Hongwei Zhang
Yahoo Japan Corporation
Tokyo, Japan
hzhang@yahoo-corp.jp

Aya Iwamoto
Yahoo Japan Corporation
Tokyo, Japan
ayiwamot@yahoo-corp.jp

Fumihiko Takahashi
Yahoo Japan Corporation
Tokyo, Japan
ftakahas@yahoo-corp.jp

ABSTRACT

In eCommerce websites, products are annotated with various metadata such as a category by human sellers. Automatic item categorization is useful to reduce this cost and have been well researched. This paper describes how we won the 2nd place (weighted F1 at Stage 1 and Stage 2 are 0.8421 and 0.8399) at SIGIR eCom DataChallenge 2018, whose goal is to predict each product's category by its title. We formulate the task as a simple classification problem of all leaf categories in a given dataset. The key features of our methods are combining of Convolutional Neural Network and Bidirectional LSTM and using ad-hoc features from an external dataset (i.e. not given in this contest). An error analysis is also employed and some cases which are hard to predict accurately are revealed.

KEYWORDS

Convolutional Neural Network, Bidirectional LSTM, External dataset

ACM Reference Format:
Shogo D. Suzuki, Yohei Iseki, Hiroaki Shiino, Hongwei Zhang, Aya Iwamoto, and Fumihiko Takahashi. 2018. Convolutional Neural Network and Bidirectional LSTM Based Taxonomy Classification Using External Dataset at SIGIR eCom Data Challenge. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR 2018 eCom Data Challenge),* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

In eCommerce websites, products are registered with metadata (e.g. title, category, etc.) by human sellers. Annotating products with those metadata is hard job, and therefore automatic predictions of metadata can reduce the cost[2]. In recent years, a number of studies of automatic item categorization in eCommerce have been made[1, 2, 8–10]. Pradipto *et al.*[1] reported that products are often categorized incorrectly because product taxonomies are large. Furthermore, if two different sellers annotate a same product with a title, the result should be different. Those difficulties cause a noisy dataset, and therefore automatic item categorization is difficult task.

At this challenge of SIGIR eCom DataChallenge 2018, Rakuten Institute of Technology provides train and test datasets. The train dataset is composed of product titles and category ID paths, and the test dataset contains only product titles. The goal of participants is to predict the category ID path for each product title in the test dataset. This challenge is more difficult than the previous item categorization problems for following two reasons: (1) The metadata of products is only title and any other information (e.g. price, image, etc.) is not included. (2) The dataset contains not category "name" paths, but category "ID" paths. This causes difficulty in using prior knowledge of each category.

This paper describes how we won the 2nd place (weighted F1 at Stage 1 and Stage 2 are 0.8421 and 0.8399) at SIGIR eCom DataChallenge 2018. We formulate the task as a simple classification problem of all leaf categories in the given dataset. The key features of our methods are following two parts: (1) Convolutional Neural Network and Bidirectional LSTM are used together. This technique may be useful because two models are different in structure. (2) Amazon Product Data[3, 6], which contains product reviews and metadata from Amazon, is used to generate ad-hoc features. The products in the dataset given by this contest do not have metadata as described above. Thus, it is useful to incorporate the metadata from the external dataset.

In the rest of the paper, the detail of our system is described in Section 2. Section 3 describes error analysis of our model. Finally, we present the conclusion in Section 4.

2 METHODS

An overview of our system is given in Figure1. A product title is fed to the system and a category for the product is predicted by following procedures: (1) The product title is split into words and they are normalized. (2) Each word is converted into an embedding vector. (3) The embedding vectors are input into "Multi-kernel CNN module" and "BiLSTM module". Each module outputs a flattened vector. (4) Ad-hoc features are generated for each word. Each ad-hoc feature is fed into a multi layer perceptron and a flattened vector is

SIGIR 2018 eCom Data Challenge, July 2018, Ann Arbor, Michigan, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

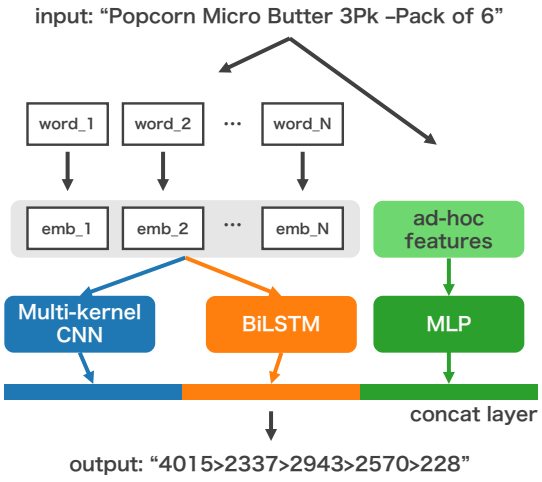

Figure 1: An overview of our system.

gotten. (5) Three vectors from step (3) and (4) are concatenated into a flattened vector and passed into a last fully connected layer. Probabilities of all leaf categories are output from the fully connected layer.

In the following of this section, we describe the detail of those procedures.

2.1 Preprocessing of a product title

First, the input sequence (i.e. product title) is split into some words by a space character. Then, symbol characters (e.g. %, #, etc.) in each word are removed. Finally, each of the words is converted into lowercase.

2.2 Generating embedding vectors

We used word2vec implemented by gensim[7] to generate skip-thoughts embedding vectors. The setting of word2vec is as follows.

- used all words appearing in train and test dataset
- window size is 7
- hierarchical softmax is used for model training
- negative sample size is 5
- embedding vector size is 512

In general, a word embedding is used in the area of natural language processing. In addition to this, we generated other embeddings as follows.

- pos tags
- stemmed word
- lemma of a word
- hypernym of a word

These embeddings are useful for unknown words in the test dataset.

2.3 Training Modules

We used Convolutional Neural Network with multiple kernels (Multi-kernel CNN)[4] and Bidirectional LSTM with Soft Attention[5] for training modules.

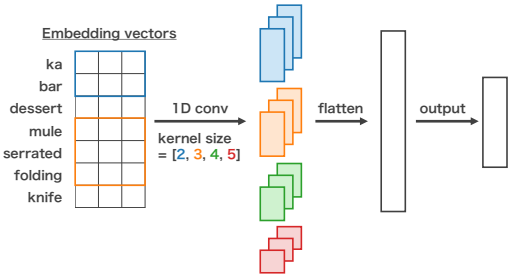

Figure 2: An overview Multi-kernel CNN module.

2.3.1 Multi-kernel CNN. Y. Kim[4] proposed a Convolutional Neural Network based method for sentence classification problem. We adopted the idea to predict categories and call this module “Multi-kernel CNN”. An overview of “Multi-kernel CNN” module is given in Figure 2.

The input of this module is embedding vectors described in Section 2.2 and the output is a vector whose elements correspond to probabilities of each leaf category. First, the input is passed into one-dimensional convolutional layers and the outputs are feature maps. We used multiple convolutional layers different in kernel size (e.g. 2, 3, 4 and 5). Next, feature maps are flattened into a vector and it is passed into a last fully connected layer.

2.3.2 Bidirectional LSTM with Soft Attention. In recent years, Recurrent Neural Networks have been used in the area of natural language processing. In the field of neural machine translation, it is reported that attention mechanism is effective technique[5].

We employed bidirectional LSTM with Soft Attention[5] to predict categories. We note that an attention layer of “sequence to sequence” model accepts what LSTM layers output from both of input-side and output-side sequences; however, this model is “sequence to label” and only output from input-side LSTM layers is accepted. An overview of “Bidirectional LSTM” module is given in Figure 3.

The input and the output is same as “Multi-kernel CNN” module. First, the input is passed into a Bidirectional LSTM layer and the output is encoded sequence. Then, the encoded sequence is passed into a soft attention layer and the output is probability distribution over all leaf categories.

2.4 Ad-hoc features

Ad-hoc features such as length of a title are useful to improve the accuracy. The list of ad-hoc features is as follows.

- title length
- uppercase rate
- alphabet/non-alphabet/digits rate count
- space character rate/count
- max length of words
- unique word rate
- a number of filtered symbols
- a number of words
- histogram of word length
- histogram of pos tags

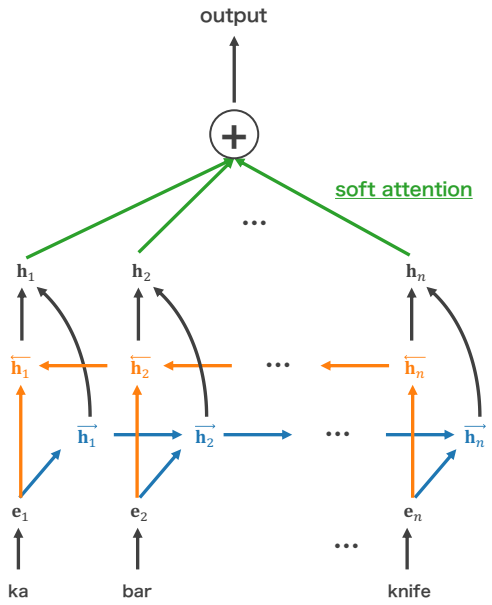

Figure 3: An overview of Bidirectional LSTM module.

In addition to ad-hoc features as below, metadata of products from an external dataset, Amazon Product Data[3, 6], which contains product reviews and metadata such as categories and prices, is used as ad-hoc features.

Figure 4 shows the procedure of generating the ad-hoc features from Amazon Product Data. (1) Training a model whose input is a title and output is a category with Amazon Product Data. (2) The product titles of the datasets given by this contest and Amazon Product Data are fed into the model of (1) and embedding vectors are gotten from an embedding layer. (3) Prices and categories of 20 products in Amazon Product Data which are similar to each product in the dataset given by this contest are fetched. The similarity of two products is the euclidean distance between two embedding vectors from (2).

2.5 Concat vectors from previous modules

We get three flattened vectors from "Multi-kernel CNN", "Bidirectional LSTM" and the multi layer perceptron with ad-hoc features. In this part, these vectors are concatenated into a flattened vector and it is passed into a fully connected layer whose output is as same as the previous module (e.g. probabilities of each leaf category).

2.6 Over sampling on "small" categories

In training, we oversampled data by shuffling words in titles for categories which have less than 50 products. It solves a problem of an imbalance of categories.

3 ERROR ANALYSIS

We split the train dataset by this contest into two parts: train and validation part to check a performance of the proposed system.

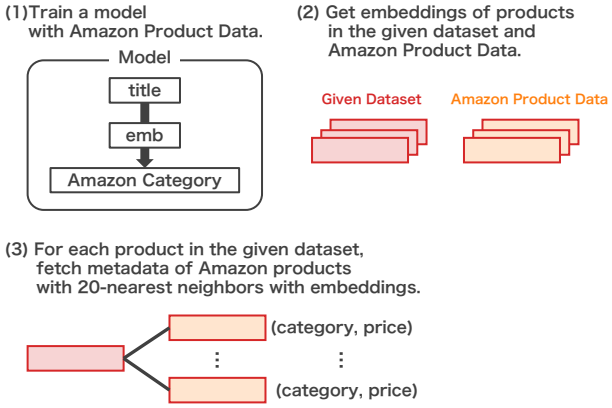

Figure 4: Generating ad-hoc features from Amazon Product Data.

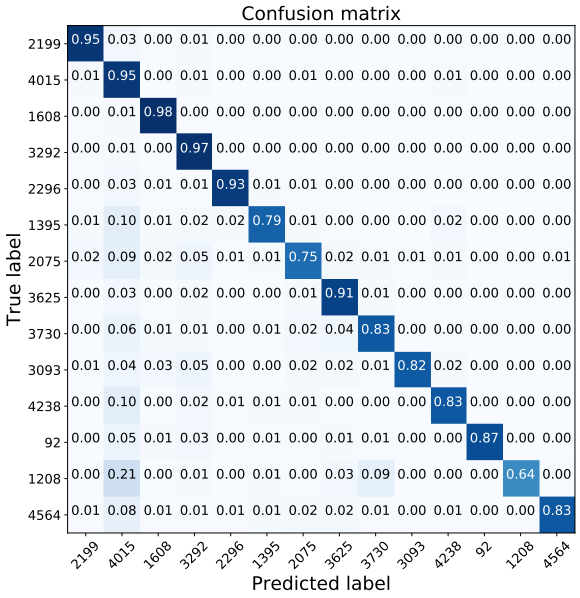

Figure 5: Confusion matrix of top level category prediction.

3.1 Top level category prediction

First, we show the accuracy of top level category prediction. Figure 5 shows the confusion matrix of top level category prediction. It is found that products whose top level category is "1208" are often miss classified as "4015". In Table 1, examples which are miss classified "1208" as "4015" and products correspond to miss classified category are shown in Table 2. In the first and second lines in Table 1, true and predicted category are seem to be similar. More specifically, "1208>310>1629>1513>3369" and "4015>4454>473" seem to be a food category. "1208>546>4262>572"

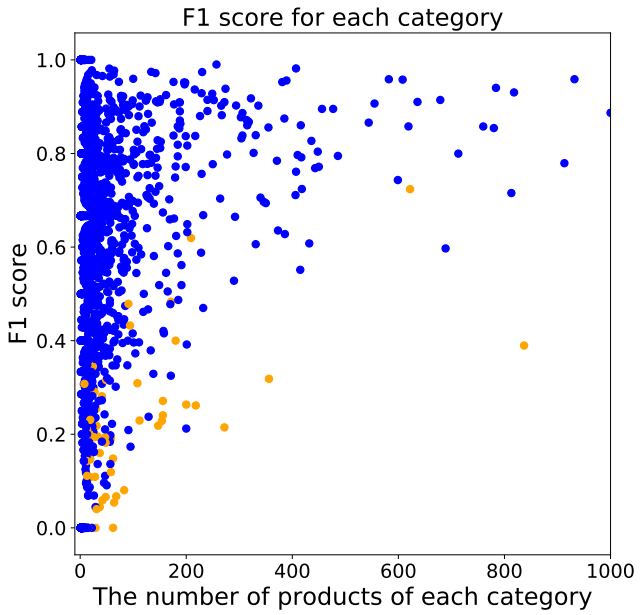

Figure 6: F1 scores for each category. The vertical line shows F1 score and the horizontal line shows the number of products in each category. The orange points show categories whose top level is "2296".

and "4015>3754>3580>1695" seem to be a DIY category. It is difficult to predict the categories of the products without a detail of the categories. For the miss classified case in the third line in Table 1, "Popcorn" is in the each title of products correspond to the predicted category "4015>2337>2943>2570>228". If products correspond to different categories have same words, it is hard to distinguish those categories.

3.2 Difficult Categories

In this sections, difficult categories are explored. Figure6 shows F1 scores for each category. It can be said that it is difficult to predict accurately for categories which few products are correspond to.

Furthermore, we focus on categories that are hard to predict in categories close to the number of products (i.e. bottom of the Figure6) and it is found that categories whose top level is "2296" are tend to bottom of the Figure6 (orange points). Table3 shows example products correspond to categories whose top level is "2296". It seems that the "2296" shows a media category, such as books and CD&DVD titles. It is found that the title of products in a media category tend not to have specific words to show its category. This causes difficulty in prediction for "2296" category.

4 CONCLUSION

In this paper, we describe how we tackle SIGIR eCom DataChallenge 2018. Our proposed model is combined Convolutional Neural Network and Bidirectional LSTM. We also used Amazon Product Data to generate ad-hoc features. In error analysis, it is found that two categories which are similar or share same words in titles are hard to distinguish. It is also found that media categories are hard to distinguish because each title of products in those does not have enough information. We believe that high prediction accuracy came from proposed deep learning models and the external dataset, but human prior knowledge for each category is useful to get better performance.

REFERENCES

[1] Pradipto Das, Yandi Xia, Aaron Levine, Giuseppe Di Fabbrizio, and Ankur Datta. 2017. Web-scale language-independent cataloging of noisy product listings for e-commerce. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Vol. 1. 969–979.

[2] Jung-Woo Ha, Hyuna Pyo, and Jeonghee Kim. 2016. Large-scale item categorization in e-commerce using multiple recurrent neural networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 107–115.

[3] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 507–517.

[4] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[5] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[6] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 43–52.

[7] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[8] Dan Shen, Jean-David Ruvini, Rajyashree Mukherjee, and Neel Sundaresan. 2012. A study of smoothing algorithms for item categorization on e-commerce sites. *Neurocomputing* 92 (2012), 54–60.

[9] Dan Shen, Jean-David Ruvini, and Badrul Sarwar. 2012. Large-scale item categorization for e-commerce. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 595–604.

[10] Yandi Xia, Aaron Levine, Pradipto Das, Giuseppe Di Fabbrizio, Keiji Shinzato, and Ankur Datta. 2017. Large-Scale Categorization of Japanese Product Titles Using Neural Attention Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Vol. 2. 663–668.

Table 1: Examples which are miss classified top level category: "1208" as "4015"

title	true category	predicted category
Wabash Valley Farms 77809 Bring Home The Bacon	1208>310>1629>1513>3369	4015>4454>473
Cartoon Growth Height Measure Chart Wall Decor Sticker DIY Wallpaper Decal	1208>546>4262>572	4015>3754>3580>1695
Popcorn Micro Butter 3Pk -Pack of 6	1208>310>397>1635>587	4015>2337>2943>2570>228

Table 2: Example products correspond to miss classified categories.

category	title
4015>4454>473	Artichoke Ccktl Marinated -Pack of 12
	AzureGreen HACTCP 2oz Activated Charcoal Powder
	Muir Glen B04579 Muir Glen Crushed Tomato In Puree - 6x104 Oz
4015>3754>3580>1695	Unique BargainsButterfly Flower Print Removable Wall Sticker Decal DIY Wallpaper Decoration
	Unique Bargains Living Room Plum Blossom Pattern Adhesive Decal Wallpaper 60 x 45cm Wall Sticker
	Brewster Home Fashions DL30463 Accents Suelita Striped Texture
4015>2337>2943>2570>228	Cuisinart Air Popcorn Maker Cuisinart Popcorn Maker
	0.75 Ounce Movie Theater Popcorn Box (Pack of 50)
	Great Northern Paducah 8oz Popcorn Popper Machine w/Cart, 8 Ounce - Black

Table 3: Example products correspond to categories whose top level is "2296".

category	title
2296>2435>1576	Rawhide Down
	Scarlet Women
	The Quartet
2296>2435>3792	Alfred 00-21113 I Will Sing - Music Book
	Nocturnes and Polonaises
	Complete Preludes and Etudes-Tableaux
2296>3597>3064	Blues Heaven
	Dowin In The Delta
	Free Beer
2296>3597>3956	Regina Belle - Believe in Me
	New Edition
	Backyard - Skillet
2296>3706>1586	Suits-Season Three
	Newlyweds-Nick and Jessica Complete 2nd and 3rd Seasons
	Defiance-s3 [dvd] [3discs] (Universal)
2296>3706>3437	Log Horizon: Season 2 - Collection 2
	Yu Yu Hakusho Season 3
	Case Closed-Season 3-S.A.V.E.