Unconstrained Production Categorization with Sequence-to-Sequence Models

Maggie Li[*]
National University of Singapore
Singapore
something@comp.nus.edu.sg

Liling Tan, Stanley Kok, Ewa Szymanska
Rakuten Institute of Technology
Singapore
{first.lastname}@rakuten.com

ABSTRACT

Product categorization is a key component to ensure e-commerce platforms to accurately retrieve the relevant products. Different from these approaches, we consider the category prediction task as a sequence generation task where we allow product categorization beyond the hierarchical definition of the full taxonomy. We build a sequence-to-sequence model to generate non-constrained product category labels.

This paper presents the results of the RIT-SG submissions for the Rakuten Data Challenge at SIGIR eCom'18 using attentional sequence-to-sequence model. The goal of the challenge is to predict the product category given the e-commerce product title.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; • **Applied computing → Electronic commerce**;

KEYWORDS

Text Classification, Sequence-to-Sequence

1 INTRODUCTION

Product categorization is a key component to ensure e-commerce platforms accurately retrieve the relevant products[9]. E-commerce sites uses hierarchical taxonomies to organized products from generic to specific classes. The taxonomies also allows easy detection of similar products that is used to power product recommendation and duplicate removals engines on e-commerce sites [14, 16]. Although merchants are encouraged to manually categorize their products when they post them on the platforms, the process is labor-intensive and leads to inconsistent categories for similar products. [3, 10]

Previous approaches to e-commerce product categorization focused on mapping the product information (titles, descriptions, images, etc.) to the specific categories based on the existing labels from the training data.

Instead of imposing the hard boundaries inherited from higher level categories, we allow the cross-pollination of sub-categories beyond the pre-defined hierarchy. For example, our Seq2Seq model was able to generate a category that was not pre-defined in training data when it assigned *"Cover Girl - Proctor Outlast Stay Luminous Foundation Cls Tan"* the `3625>2644>2805>1615` label. For reference, table 1 shows a sample of related product titles and their respective categories from the training data that overlapped with the `3625>2644>2805>1615` label.

Category: 3625>2644>921>1615
Clean Matte Pressed Powder - # 545 Warm Beige by CoverGirl for Women - 0.35 oz Powder
Laura Mercier Silk Creme Oil Free Photo Edition Foundation - Sand Beige 1oz
Aqva Pour Homme After Shave Balm (Tube) - 100ml/3.4oz
IMAN Second to None Luminous Foundation, Clay 5 .35 oz (10 g)
Category: 3625>2644>2805>3870>1102
Makeup Blender, Assorted Colors, 10 Count
Category: 3625>2644>2805>2522
SHANY Detox Professional Brush Cleanser - Instant dry - Refill - 16oz
Category: 3625>2644>2805>3870>2697
8Pcs Make Up Cosmetic Brushes Set Powder Foundation Eyeshadow Lip Brush Tool Kit
Category: 3625>2644>2805>3870>4627
Studded Couture - 12 Piece Brush Set

Table 1: Product Titles and Categories in the Training Data

2 SEQUENCE-TO-SEQUENCE LEARNING

The most common Seq2Seq models belong to the encoder-decoder family where the source sequence is encoded as a fixed-length vector and then fed to a decoder steps which will step through to generate the predicted output sequence one symbol at a time until an end-of-sequence (EOS) symbol is generated. The encoder and decoder is jointly trained to maximize the probability of generating the correct output sequence given its input [4, 5, 8, 11].

[*] This is the corresponding author

Conference'17, July 2017, Washington, DC, USA

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

Top-level Categories	Count	(%)	Largest Sub-category	(%)
4015	268,295	0.3353	4015>2337>1458>40	0.031851
3292	200,945	0.2511	3292>3581>3145>2201	0.037682
2199	96,714	0.1208	2199>4592>12	0.087393
1608	85,554	0.1069	1608>4269>1667>4910	0.013727
3625	29,557	0.0369	3625>4399>1598>3903	0.021400
2296	28,412	0.0355	2296>3597>689	0.004927
4238	23,529	0.0294	4238>2240>4187	0.001985
2075	20,086	0.0251	2075>4764>272	0.004962
1395	18,847	0.0235	1395>2736>4447>1477	0.004720
92	8172	0.0102	92	0.010215
3730	8113	0.0101	3730>1887>3044>4882	0.003978
4564	5648	0.0070	4564>1265>1706>1158>2064	0.001281
3093	5098	0.0063	3093>4104>2151	0.001907
1208	1030	0.0012	1208>546>4262>572	0.000195

Table 2: Distribution of First Level Categories and the Most Common Label in Each First Level Categories

Simple encoder-decoder performance deteriorates when translating long input sequences; the single fixed-size encoded vector is simply not expressive enough to encapsulate that much information. Bahdanau (2014) proposed the attention mechanism that learns an implicit alignment between the input and output sequences. Before the decoder generates a item, it first aligns for a set of positions in the source sequence with the most relevant information.[1] The model then predicts the target item based on the context vectors of these relevant positions and the history of generated items.

3 DATASET CHARACTERISTICS

The Rakuten Data Challenge (RDC) dataset consists of 1 million product titles and anonymized hierarchical category labels. The data was split 80-10 into training and testing set. For the competition, the test labels were kept unknown until the end of the competition. Like most e-commerce product categorization data [2, 6, 15], the distribution of the products' 14 top-level categories are non-uniformly distributed. From the training set, there are 3000 unique categories and the largest category (2199>4592>12) contains ~69,000 product titles that made up 8.7% of the 800,000 product titles from the training set.

It is common for E-commerce text datasets to be inherently noisy; recent related works on product categorization had dedicated approach to address the noise through a combination of feature engineering and classifier ensembles.[3, 10] From the RDC dataset, we checked for common noise signatures by checking for product titles that contains characters beyond the printable ASCII range (0x20 to 0x7E). Figure 1 shows the list of characters outside the range, the left side shows the number of product titles that contains one or more of the characters on the right.[1]

```
1       ['\x9d', '\x9f', ', ', '\xad', '¨', '⅜', '⅝',
         'i', ¦', 'm²', 'µ', 'œ', !̥', 'ђ', 'њ', '´',
         'ù', '\x96', '', ' (', ') ', 'ō', '×', '£',
         '↑', '\x81', '| ']

2-10    ['ú', '\x94', '\x7f', 'ʙ', 'ē', 'ç', 'ū', '-',
         '—', '⅛', '•', 'ï', '¿', '·', 'ê', 'º', '¦',
         'ä', 'ñ', '²', '├', '⌐']

10-50   ['ñ', '²', '├', '⌐', 'ö', '-', '¼', '†', 'ž',
         '±', '¡', '…', 'í', 'μ', '„', 'ó', 'å', 'á',
         'æ', '¾', '\x99', 'š'

>50     ['¬', '€', ',', 'ƒ', 'ë', '¢', 'ã', 'â', '\xa0', '�']
```

Figure 1: Lists of Characters not in Printable ASCII Range

Submissions	Configurations	P	R	F
Submission 1	Encoder-Decoder with Attention (EDA)	0.77	0.78	0.77

Table 3: Precision, Recall, F1 Scores on Held-out Test Set

4 EXPERIMENTS

Without explicit tuning, we trained a single-layer attentional encoder-decoder network for two hours using the Marian toolkit[7] with the following hyperparameters

- **Batch size:** 5000
- **Dropout:** 0.1 (Embeddings and RNN layers)
- **Beam Size:** 6
- **Epochs:** 7

5 RESULTS

The results on Table 3 showed that without explicit pre-processing and tuning, a baseline Seq2Seq model can achieve competitive results scoring 0.77 on weighted F1-score.

[1] It is worth noting that while the characters >50 is small, most of them appear in 50-200+ product titles. And the last second character in the list non-breaking spaces `xa0` and the last replacement character appears in 643 and 766 product titles respectively. Usually, these are breadcrumbs of the HTML to Unicode conversion.[12, 13]

6 CONCLUSION

By framing the product categorization task as a sequence generation task, we trained a attentional sequence-to-sequence model to generate non-constrained product that is not limited to the supervised categories from the training dataset. We achieved an F1-score of 0.77 in the Rakuten Data Challenge at SIGIR eCom’18 .

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Ali Cevahir and Koji Murakami. 2016. Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 525–535.

[3] Jianfu Chen and David Warren. 2013. Cost-sensitive Learning for Large-scale Hierarchical Classification. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management (CIKM ’13)*.

[4] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics.

[5] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

[6] 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.

[7] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, AndrÃl’ F. T. Martins, and Alexandra Birch. [n. d.]. Marian: Fast Neural Machine Translation in C++. *arXiv preprint arXiv:1804.00344* ([n. d.]). https://arxiv.org/abs/1804.00344

[8] Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

[9] Bhargav Kanagal, Amr Ahmed, Sandeep Pandey, Vanja Josifovski, Jeff Yuan, and Lluis Garcia-Pueyo. 2012. Supercharging Recommender Systems Using Taxonomies for Learning User Purchase Behavior. In *Proceedings of VLDB Endowment*.

[10] Zornitsa Kozareva. [n. d.]. Everyone Likes Shopping! Multi-class Product Categorization for e-Commerce. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, year = 2015*.

[11] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*.

[12] Liling Tan and Francis Bond. 2011. Building and Annotating the Linguistically Diverse NTU-MC (NTU-Multilingual Corpus). In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*.

[13] Liling Tan, Marcos Zampieri, Nikola Ljubesic, and Jorg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*.

[14] Li-Tung Weng, Yue Xu, Yuefen Li, and Richi Nayak. 2008. Exploiting Item Taxonomy for Solving Cold-Start Problem in Recommendation Making. In *2008 20th IEEE International Conference on Tools with Artificial Intelligence*.

[15] 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*. Association for Computational Linguistics.

[16] Cai-Nicolas Ziegler, Georg Lausen, and Lars Schmidt-Thieme. 2004. Taxonomy-driven computation of product recommendations. In *CIKM*.