Toy Matching in E-commerce

Kai Lin Erzhen Hu Yinzhu Jin Fuxiao Liu

University of Virginia, Department of Computer Science

Abstract

The purpose of this project to detecting categories from the toy product description. The data set is an Amazon (toy) data set with manufacturer specific model, names and description of children's toys.
The expected outcome includes labeling scheme and use CRF and bi-LSTM to measure the performance of our category extraction. As for the annotation, we are doing the labeling and used a product description decomposed into a sequence of tokens labeled with BIO encoding, and the output of learning algorithm on a product description would a sequence of labels.

We tried three methods, the first one is using pre-trained word embedding to improve the classification performance, here we tried several models like svm, logistic regression, and the linear svm get the highest classification accuracy, there' s too many categories in such a small dataset
The expected outcome includes labeling scheme and use CRF and bi-LSTM to measure the performance of our category extraction.

Introduction

In 2017, Americans buy at least one in every three items online. And the most common way most shoppers try to find what they're looking for is search. There are mainly three search query types: Product type search, Feature search (Cheap red evening gown), Subjective search (High-quality sofa). However, even the top 50 grossing US ecommerce websites don't do a great job supporting some common types of search queries, resulting in irrelevant products or zero results.

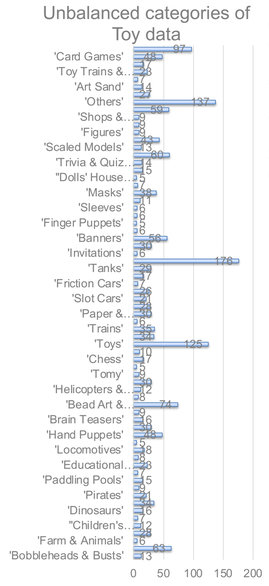


Figure 2. categories

Groups of identical items sold by different sellers are formed and on the other hand, we want to be precise in the creation of categories to ensure they obtain identical products sold by different sellers. So if the categories is formed, it will represent exactly one item page with a single title, description and price. But if the groups is not homogeneous, then it will be showed on a page that it is not belongs to.
To make the process specific, let x be a product description and let (x 1 , x 2 , .., x n) be a particular tokenization x t of x. Given a category η, extraction is the process of discovering the functions:

Dataset	Sample number
training set	5555
validation set	1864
test set	1893
total	9312

Table 1: Dataset Statistics

	Unnamed: 0	category	description	product_name
0	0	Bobbleheads & Busts	Your favorite marvel super heroes available in...	Funko Thing Pop Vinyl Figure Bobble Head
1	1	Bobbleheads & Busts	Product Description Fred "Freddy" Krueger is t...	A Nightmare on Elm Street - Freddy Krueger Bob...
2	2	Bobbleheads & Busts	Product Description For the first time ever, A...	Assassin's Creed Brotherhood Ezio Headknocker
3	3	Bobbleheads & Busts	Product Description Dark Knight Rises: Batman ...	Dark Knight Rises: Batman Head Knockers
4	4	Bobbleheads & Busts	The Hobbit - An Unexpected Journey - Bilbo Bag...	The Hobbit Bilbo WackY Wobblers/Wackelkopf [Ge...

Figure 1. Data set of our project

Literature review

The Name entity recognition(NER) task can be used beyond identifying people, location. Many recent studies (Putthividhya and Hu,2011;More,2016) focused on extracting product attributes from online E-Commerce data. One of the popular generative model based classifiers for NER is Hidden Markov Model, which captures the temporal states by modeling transitions over time.However, methods like SVM for NER have been shown to outperform generative model.
Moreover, Conditional Random Fields(Feng and McCallum 2004) has been proposed for sequence labeling problem for NER. Additionally, Majumder,et al.(2018) demonstrated the potential of Bidirectional LSTM-CRF and Bidirectional LSTMs in E-commerce domainwhich shows the efficacy of these deep recurrent models over previous machine learning benchmarks. To solve the existing grammatical structure of E-commerce data in both product title and product description, some studies(Putthividhya and Hu,2011) used the boostrapped Named Entity Recognition to identify new brands corresponding to spelling variants and some typographical errors.
There are still challenges in product descriptions that, these descriptions may contain lots of named entities unrelated to the product instead of product title. The number of different attributes that are used by a general retailer to describe products

Methods

We want to show here how we define the metrics used to measure the performance of the category extraction algorithm.

- **Method1:** Word embedding using pretrained glove 50-d vectors with classification methods(logistic regression, svm, NB)
- **Method2:** Here we implement the CRF model with extracted features(word parts, POS tags, lower/title/upper flags, features of nearby words), the NER tag (B-category name, I-category name, O)+linear-chain CRF, we got the state features and transition features from the CRF model, we set 3-fold cross validation to get optimized results.
- **Method3:** Bi-LSTM: word embedding using pretrained glove 50-d vectors, tag embedding initialized with one-hot representation, then learned along the training.
- For the pre preprocessing part, we have extracted the last hierarchy of the Amazon categories' hierarchical structure and marked 169 categories with frequency less than 25 as a new category, "Others". we also implement train test split on this phase.

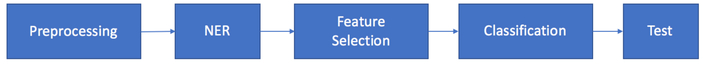


Figure 2. Overall Project Architecture

Discussion & future works

- We tried three methods, the first one is using pre-trained word embedding to improve the classification performance, here we tried several models like svm, logistic regression, and the linear svm. SVM performs better than other text classification methods after pre-trained word embedding. Conditional random fields gets a high micro average f1_score but a low macro average f1_score due to the zeros of many categories, decreasing the C1 might be a good way but influence the model quality here. Bi-directional LSTM did not perform good in this experiment.
- The main challenge for us is still that there' s too many categories in such a small dataset, also, the categories is not balanced. The reasons why the SVM performs better than other text classification methods after pre-trained word embedding is might because the sentences are not grammatically complete, i.e.
- (The Hobbit - An Unexpected Journey - Bilbo Baggins Bobble Head Figure)
- As for the future work, since it was a sub-set from Amazon dataset, it would be better if we could try on larger datasets to compare the mdoels. The absence of syntactic structure in such short pieces of text makes extracting attribute values a challenging problem.

Results

1 Text classification with pre-trained word embedding
Here, we tried word embedding using pretrained glove 50-d vectors with SVM of different kernels, logistic regression, NB, etc. the results showed that linear SVM get the best result

	Precision	Recall	F1score	time
svm (kernel='linear', c=0.1)	0.309	0.26131479	0.27502532	0.2s

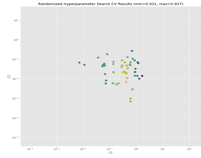
2 NER with conditional random fields(3fold cross validation)
We tried linear-chain CRF here for the pre-defined categories,

		Precision	Recall	F1score	time
CRF with cv ('c1': 0.42139673260345256, 'c2': 0.0067941467019329285 (best CV score: 0.9367260174060519)	Micro average	0.954	0.955	0.955	1h 32min 3s(150 fits)
	Macro average	0.247	0.154	0.176	
CRF with cv ('c1': 0.00001, 'c2': 0.0067941467019329285 (decreasing c1 to get l1 regularization)	Micro average	0.998	0.998	0.998	4min 20s
	Macro average	0.319	0.205	0.236	

Table 1. transition and state features from Conditional Random fields

From \ To	O	B-Chess	B-Hand Puppets
O	2.572	0.029	-0.014
B-Chess	0.826	-0.135	0.0
B-Hand Puppets	0.074	0.0	-1.335

The model learns that token "champion" ,"woods", "recognized" ,"supplied", "officially" is likely to be at the beginning of the category "chess"; wry could be at the beginning of category Hand Puppets



y=O top features		y=B-Chess top features		y=B-Hand Puppets top features	
Weight[?]	Feature	Weight[?]	Feature	Weight[?]	Feature
+7.708	postag:NNS	+4.051	word.lower():chess	+4.369	+1:word.lower():wry
+6.323	postag:NNP	+3.019	+1:word.lower():chess	+3.337	-1:word.lower():art
+6.006	bias	+2.974	-1:word.lower():champion	+3.119	-1:word.lower():over
+5.402	postag:NNPS	+2.452	-1:word.lower():woods	+2.970	-1:word.lower():take
+4.615	-1:postag:NN	+2.413	-1:word.lower():?	+2.827	-1:word.lower():giraffe
… 876 more positive …		+2.147	-1:word.lower():chess	+2.811	-1:word.lower():curiosity
… 873 more negative …		+1.978	+1:postag:VBN	+2.791	+1:word.lower():art
-2.939	-1:word.lower():chapter	+1.936	+1:word.lower():recognized	+2.773	+1:word.lower():monkey
-2.967	+1:word.lower():beaten	+1.775	+1:word.lower():men	+2.757	word.lower():dimension
-2.968	+1:word.lower():d6	+1.745	word[-2:]:ss	+2.719	-1:word.lower():dad
-3.852	word.lower():spinning	… 69 more positive …		… 153 more positive …	
-4.510	postag[:2]:NN	… 5 more negative …		… 14 more negative …	

4.368741	B-Hand Puppets	+1:word.lower():wry
4.164948	B-Packs & Sets	+1:word.lower():heartgold
3.996605	B-Art Sand	+1:word.lower():magic
3.985012	I-Dolls	+1:word.lower():soft
3.977346	B-Balloons	+1:word.lower():balloons
3.857763	B-Card Games	+1:word.lower():facet
3.810840	B-Slot Cars	-1:word.lower():mm
3.704961	B-Dice & Dice Games	+1:word.lower():d6
3.391243	B-Card Games	-1:word.lower():u.s

3 Bi-LSTM with pretrained glove 50d vectors
Majumder,et al.(2018) demonstrated the potential of Bidirectional LSTM-CRF and Bidirectional LSTMs in E-commerce domain which shows the efficacy of these deep recurrent models over previous machine learning benchmarks, here we tried Bi-LSTM but the preliminary results is not that good.

	Precision	Recall	F1score	time
Bi-lstm	0.003	0.014	0.003	

Contact

Kai Lin: kl5ev@virginia.edu

References

[1] Putthividhya, D.P. and Hu, J., 2011, July. Bootstrapped named entity recognition for product attribute extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1557-1567). \Association for Computational Linguistics.
[2] More, A., 2016. Attribute extraction from product titles in ecommerce. arXiv preprint arXiv:1608.04670
[3] Peng, F., Feng, F. and McCallum, A., 2004, August. Chinese segmentation and new word detection using conditional random fields. In Proceedings of the 20th international conference on Computational Linguistics (p.562). Association for Computational Linguistics.
[4] Majumder, B.P., Subramanian, A., Krishnan, A., Gandhi, S. and More, A., 2018. Deep Recurrent Neural Networks for Product Attribute Extraction in eCommerce. arXiv preprint arXiv:1803.11284