Attribute extraction from eCommerce product descriptions

Mikhail Sidorov (SUID: msidorov)

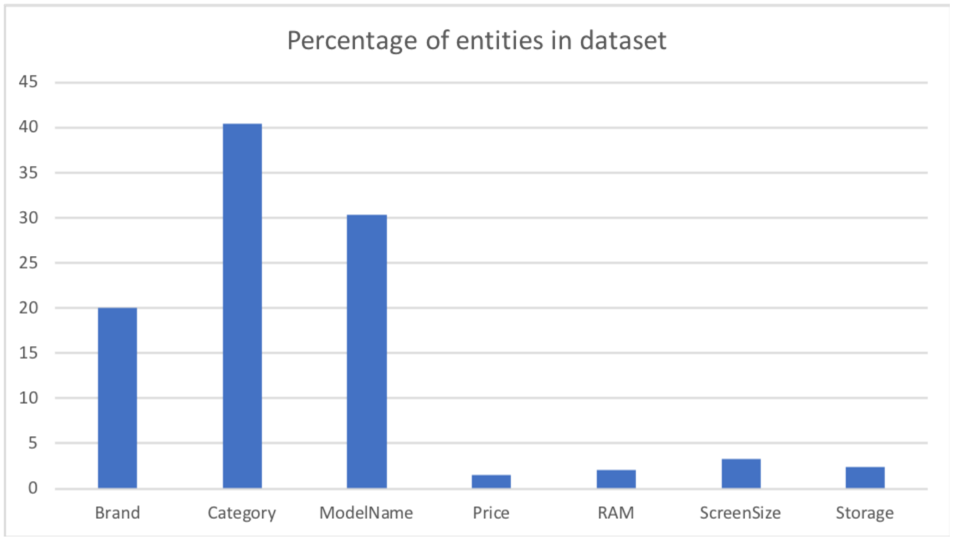
1. Introduction

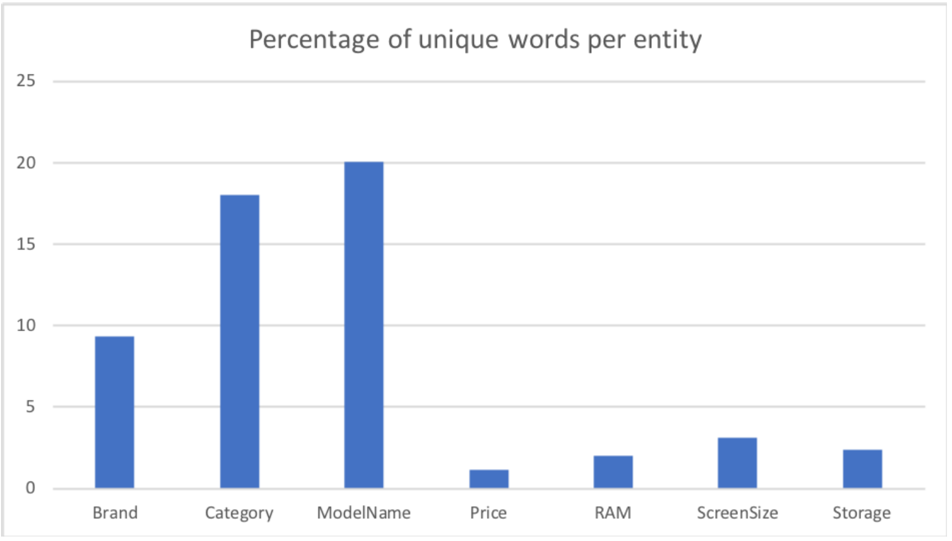
This project presents an implementation of named entity extraction for detecting attributes in the description of eCommerce products. This problem is very important for eCommerce search and catalog building systems. Effective named entity extraction could significantly improve quality of search results in eCommerce retail system and so the experience of customers. Because description of products is provided in plain text form without any structuring, this is also very challenging problem. Using as an example BestBuy eCommerce NER dataset we demonstrate the technology which includes feature extraction pipeline and training the model to recognize Brands, ModelNames, Price and other attributes from the product description. For tagging eCommerce product description we compare classification approach (SVM/GBT) with the probabilistic models HMM/MEMM/CRF.

2. Dataset description

We use BestBuy eCommerce dataset provided for NER problem. Exposed data set has about 4000 records annotated by experts.

Apple	watch	series	3	42mm	from	$339
Brand	Category	ModelName	ModelName	ScreenSize	None	Price





3. Feature extraction pipeline

Feature extraction pipeline responsible for flexible extraction approach suggested [1]

Document Corpus
Feature Extraction

Feature	Description
w_0	Represent token in current position (one hot)
w_{-1};w_0	Bigram (one hot)
w_0 is a number	1 if token has only digits
w_{-1} == (and)	1 if previous token is and
Count of chars in w_0	Length of the token
Token position	Position of token in the document
w_0 is uppercase	1is all letters are CAPS

References

[1] Ajinkya More. Attribute extraction from product titles in ecommerce. *WalmartLabs*, 2016.

[2] Sivaji Bandyopadhyay Asif Ekbal. Named entity recognition using support vector machine: A language independent approach. 2010.

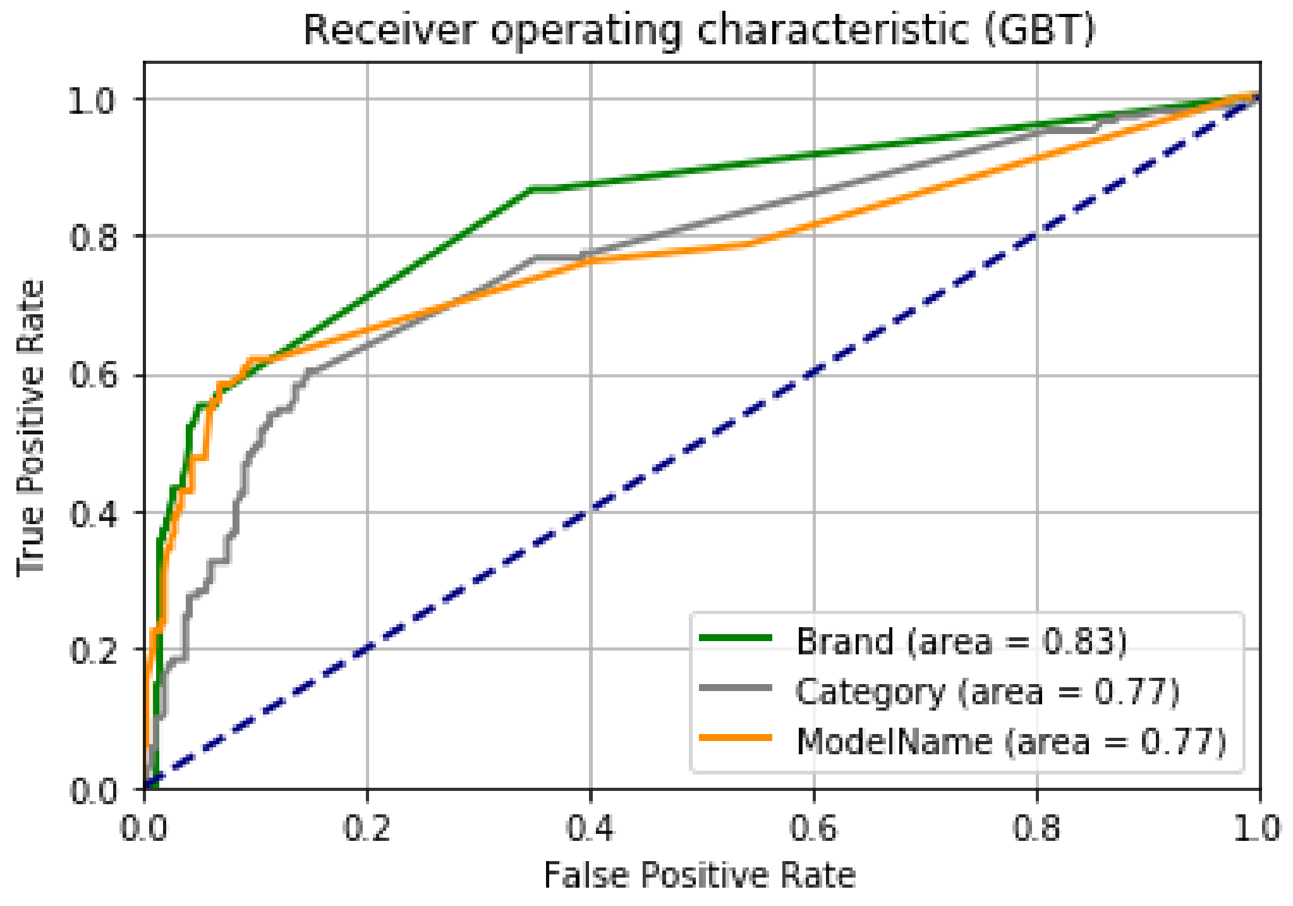
[3] Dan Jurafsky and James H. Martin. Speech and language processing (3rd ed. draft). 2018.

4. NER using Classification

Multinomial classification approach to classify the tokens which are represented as feature vectors. Classification has been implemented using sklearn library SVM [2] and GradientBoostingClassifier. For result esimation we use *Precission*, *Recall* and F_1 metrics. The definition are:

$$Precision = \frac{TP}{TP + FP};\ Recall = \frac{TP}{TP + FN};\ F_1 = \frac{Precision \times Recall}{Precision + Recall}$$

		SVM			GBT		
		precision	recall	f1-score	precision	recall	f1-score
Training	Brand	0.900	0.750	0.820	0.870	0.830	0.850
	Category	0.900	0.800	0.840	0.830	0.840	0.840
	Model	0.930	0.780	0.840	0.860	0.800	0.830
	Total	0.911	0.782	0.835	0.849	0.823	0.838
Test	Brand	0.895	0.486	0.630	0.852	0.605	0.708
	Category	0.667	0.531	0.591	0.604	0.527	0.563
	Model	0.800	0.431	0.560	0.622	0.426	0.505
	Total	0.768	0.482	0.587	0.675	0.510	0.579



5. Probabilistic model (Conditional random field)

We also used probabilistic approach and train the model using CRF [3]

$$p(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} exp\left(\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right)$$

		CRF		
		precision	recall	f1-score
Training	Brand	0.925	0.928	0.927
	Category	0.869	0.964	0.914
	Model	0.908	0.940	0.924
	Total	0.895	0.948	0.920
Test	Brand	0.611	0.512	0.557
	Category	0.616	0.770	0.684
	Model	0.491	0.571	0.528
	Total	0.583	0.661	0.616

6. Conclusions

Following methods are using for NER:

- rule-based
- classification-based
- probabilistic (HMM/MEMM/CRF)
- based on neural networks

In current project we applied 2 of them (probabilistic and classification). CRF demonstrated better results. We didn't use gazetteers as a source for additional features. which definitely should be a strong signal. Also we had very limited data set and didn't use pretrained word embedding, which could compensate small data set. We can see that under these conditions probabilistic model demonstrated better results and less affected by overfitting. We consider to try neural architectures for NER as a continuation of this work.