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.

3. Feature extraction pipeline

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

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 estimation we use Precision, Recall and F1 metrics. The definition are:

\[ \text{Precision} = \frac{TP}{TP + FP}; \quad \text{Recall} = \frac{TP}{TP + FN}; \quad F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

5. Probabilistic model (Conditional random field)

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

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

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.

References