EVALUATION OF NAMED ENTITY RECOGNITION FOR THE GERMAN E-COMMERCE DOMAIN

Sergej Denisov and Frederik S. Bäumer
Bielefeld University of Applied Sciences, Bielefeld, Germany

ABSTRACT
Large online marketplaces offer search engines as an important navigation aid for their customers to navigate through the enormous number of different products and merchants. The quality of the search results depends to a large extent on the product information provided by the retailers. One way to improve search quality is to perform linguistic enhancement of the product data. For this, we use Named Entity Recognition to identify specific e-commerce entity types and give them higher weighting in the search. Typical entity types for the e-commerce domain are products, brands, and various product attributes. Recognition of e-commerce entity types for the German language remains a challenge due to the limited availability of existing resources and linguistic complexity. We address this challenge by acquiring data from two online e-commerce marketplaces to create six NER datasets based on German product titles and descriptions. Across these datasets, we evaluate the NER performance of the state-of-the-art models mBERT, GermanBERT, and XLM-RoBERTa. As a result, the XLM-RoBERTa model archived the best performance with an F1 score of 0.8611 averaged over all datasets.

KEYWORDS
Information Retrieval, Transformer, Named Entity Recognition, E-Commerce

1. INTRODUCTION
Modern artificial intelligence models based on the transformer architecture have proven to achieve state-of-the-art results in downstream NLP tasks (Lothritz et al. 2020). Furthermore, recently several multilingual models have been released that improve cross-lingual language understanding (Conneau et al. 2020) and can address research questions such as domain-specific Named Entity Recognition (NER) (Tjong Kim Sang and De Meulder 2003) or Named Entity Recognition in Query (NERQ) (Du et al. 2010) for different languages.

NER is a subtask of Information Extraction (IE) (Carstensen et al. 2009) and deals with the Identification and Classification of named entities in texts. While NER usually focuses on more extended texts, NERQ is primarily for web search applications and focuses on queries, which tend to be short. This work focuses on the e-commerce domain, where it is essential to extract named entities from the unstructured product information data and user search queries. As the products offered on online marketplaces come from different merchants, the quality of product descriptions varies significantly. While professional merchants usually provide detailed and structured descriptions of the products they offer, most sellers provide only unstructured descriptions (Joshi et al. 2015: 1). Because of this and the vast number of products on offer, marketplace vendors must ensure that customers get the best possible results for their search query. Our proposed approach can classify tokens from search queries, product titles, and descriptions into predefined e-commerce entities, e.g., brand and product. Enriching data with entities can improve understanding of users’ search intentions and therefore provide better search results. As a result, a more pleasant shopping experience leads to higher customer loyalty.

Our research approach, which we briefly present in this short paper, consists of two separate tasks. For the first task, we acquire product titles and descriptions in the German language from online marketplaces. Following, we improve the data quality by applying various preprocessing steps to the HTML documents. Next, the preprocessed data is analyzed to verify the quality of the data and guarantee diversification within the categories. Finally, for the annotation with the Prodi.gy framework (Montani and Honnibal 2018), we
take subsamples of data to create six German datasets for product titles and descriptions. The second task consists of evaluating models based on the transformer architecture, such as BERT (Devlin et al. 2018), and XLM-RoBERTa (Conneau et al. 2020). Finally, we use the created datasets to fine-tune the models and apply the Weights & Biases library (Biewald 2020) for hyperparameter search.

2. DATASETS AND TRAINING

Since there are no publicly available German e-commerce datasets, we had to collect data from various online marketplaces. Therefore, we acquired data from Amazon and Online-Shop to create our own datasets. This section provides insight into the data preprocessing and the distribution of entities. Furthermore, we describe our method for fine-tuning the transformer-based models with the six resulting German datasets.

2.1 Datasets

The Scrapy framework was used for the data acquisition. We collected a total of 1,871,200 product records for 74 different categories from Amazon and a total of 16,159 product records from the Online-Shop. Merchants on Amazon publish the content according to their quality standards, while the content in the Online-Shop is professionally maintained. Amazon data was collected between October and December 2019, and Online-Shop data in April 2020. We applied several heuristics to preprocess the captured data for the final annotation process. These include extracting text data from HTML, deduplicating text, removing unwanted characters, discarding product titles with less than one token, and product descriptions with less than two sentences. We were additionally filtering out product descriptions and titles that were not classified as German by the langdetect framework with a probability of more than 0.999. Finally, we use scikit-learn for calculating pairwise cosine similarity to find the most diverse product titles and descriptions. Then, we randomly select 1,000 documents for the different categories to create the datasets. In the annotation phase, we focused on the two different categories of computer and automotive. We selected the following e-commerce entities: “Brand”, “Product”, “Model”, “ItemNo”, “Quantity”, “Color”, “Size” and “Attribute”. In addition, we annotated some product attributes that do not belong to the selected categories. The intention is that the datasets can be extended with more entities, e.g., the new entity “State” for used products. The tag distribution over the eight entities for each product title and description dataset is shown in Figure 1.

![Figure 1. Tag distribution for the title (left) and description (right) datasets.](image)

For a more precise overview, the B, I, L and U tags have been combined and the O tag removed

In Figure 1, it is shown that the distribution of entities is uneven. The title datasets have more entities than the description datasets. For the titles and description datasets, the most frequent entity is “Attribute”. Besides that, “Size”, “Product”, and “Model” appear frequent. The entities “Quantity”, “Color” and “ItemNo”, on the other hand, occur less frequently.
2.2 Training of the Models

Lately, many different pre-trained models based on the transformer architecture have been released. In this work, we used a CRF model (Lafferty et al. 2001), as our baseline approach and compared it with mBERT (Devlin et al. 2018), GermanBERT1, and XLM-RoBERTa. The BERT architecture consists of 12 bidirectional transformer encoder blocks, 768 hidden layers, 110 million parameters (Devlin et al. 2018: 4173) and uses the Attention mechanism (Luong et al. 2015). The original BERT model was pre-trained based on two pre-training tasks. The first task is the “Masked Language Modeling” (MLM), and the second task is the “Next Sentence Prediction” (NSP) (Devlin et al. 2018: 4174). Within the MLM task, the model predicts 15% of random masked words from a sentence. In the second task, the model gets pairs of sentences as input and predicts if the second sentence in the pair is the original document’s subsequent sentence. XLM-RoBERTa model was trained with an MLM objective like BERT on monolingual data from 100 languages and therefore improved cross-lingual language understanding (XLU) and achieves state-of-the-art performance for various languages in different tasks (Conneau et al. 2020).

We used the CRFsuite implementation for the CRF model and did not specifically optimize the hyperparameters for the CRF model (Lafferty et. al 2001), as this model served as a baseline for all further experiments.

For fine-tuning the Transformers models, we worked with the Python library Simple Transformers (Rajapakse 2019). It has been observed that for Transformers models, large datasets (e.g., more than 100,000 labeled training examples) are much less sensitive to the choice of hyperparameters than small datasets (Devlin et al. 2018). Therefore, finding the best performing hyperparameters for our relatively small datasets is essential.

We applied the Weights & Biases Framework for hyperparameter search and experiment tracking. It supports running Sweeps for the model optimization. We conducted Bayes searches to determine optimized hyperparameters for the mBERT, GermanBERT, and XLM-RoBERTa models. After running the hyperparameter optimization on our datasets, we discovered that a batch size of 8, a learning rate of 5e-5, and 4 training epochs produced the highest F1 scores for the mBERT and GermanBERT model. For XLM-RoBERTa, the highest F1 score was achieved with the same batch size and learning rate but 10 training epochs.

3. EVALUATION AND DISCUSSION

The split for the datasets is 80/20 (training/evaluation). For measuring the performance, precision, recall, and F1 scores are used. Because of the imbalanced distribution of the entity types, we focus on the micro-averaged calculation of the scores. This calculation takes each entity type’s frequency into the count. In Table 1, the results for the four models and all datasets are shown.

---

1 Available at https://www.deepset.ai/german-bert, last accessed 2021-07-19.
As shown in Table 1, the three transformer-based models perform better than the CRF model. The transformer-based models achieve higher precision, recall, and F1 scores for each dataset than the CRF model. In particular, the results show that the XLM-RoBERTa model provides the highest F1 scores for each dataset. In addition, the transformer-based models achieve higher precision, recall, and F1 scores for each data set than the CRF model.

The conducted experiments have shown that the Online-Shop datasets achieve better results than the Amazon datasets. These results conclude that the quality of the data heavily influences the performance of the individual models. The Amazon data contains mostly unstructured and inaccurate titles and descriptions, while the Online-Shop data has a specific structure because the content is professionally maintained. Moreover, the Amazon product titles and descriptions contain several errors. The most critical ones are spelling and grammatical errors. Furthermore, there are products in the wrong category with incorrect information. Besides that, text formatting errors like missing spaces are present.

The analysis of the performance of the different entities showed that the distribution of entities within the dataset should be as balanced as possible for better results. Especially the entity types with a small number of entities performed poorly. Therefore, optimization towards more annotations for entity types with few examples is crucial. To determine if we annotated enough data and if larger datasets would perform better, we trained all models with a proportion of the Online-Shop (T) dataset. Therefore, we reduced the dataset size from 1,000 examples to 250, 500, and 750 examples. As a result, with only 250 examples of the Online-Shop title dataset, the CRF model achieves an F1 score of 0.7452, and the transformer-based models achieve an F1 score between 0.7759 and 0.8376. For 500 examples, the best result achieves XLM-RoBERTa with an F1 score of 0.9219. Moreover, using more than 500 examples of the datasets only leads to a slight increase in the F1 score.

4. CONCLUSION

In this work, we examined search optimization for online marketplaces by applying NER. This approach is essential if the search results of online marketplaces do not include any or only a few of the desired products for the search query. To address this research, we created six German e-commerce NER datasets consisting of product titles or descriptions. These datasets were essential to training the CRF model and fine-tuning the three transformer-based models. Moreover, our experiments reveal that the distribution of entities and data quality is crucial for the model results.

For the best possible NER results, the distribution of entities should be as even as possible. Therefore, the most rarely occurring entities must be supplemented with additional annotations. For our datasets, it is also possible to obtain additional entities from the “Property” entity. It would result in a more balanced
distribution of entity types. Furthermore, the created datasets confirm that online marketplaces have different quality standards for publishing product titles and descriptions. Therefore, improvement of data quality should be achieved by further preprocessing.

In summary, our experiments show that the transformer-based models achieve excellent NER performance for the German language with a small amount of data. Furthermore, XLM-RoBERTa provides the best performance for all datasets with an average F1-score of 0.8611. Consequently, XLM-RoBERTa can be used for product search engine optimization for the German e-commerce domain.

REFERENCES