Named Entity Recognition System for E-Commerce Product Titles

Trishal Gayam (1225073096), Venkata Pavan Boppudi (1226122224),
Hirthik Mathavan (1225184012), Richa Parte (1225646136)
and Awani Kendurkar (1225438149)

December 8, 2022

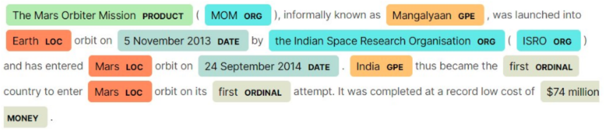
Abstract

One of the essential steps in search query comprehension is Named Entity Recognition (NER). Identification of brand and product type entities in the e-commerce space can assist a search engine in returning accurate results and, as an outcome, can provide a fun shopping experience. NER has its own challenges due to the ambiguity in data, noisy text, strong and weak entities, missing information, etc. This project also aims to address these challenges. We propose herein a **Named Entity Recognition System for E-Commerce Product Titles**. We plan to extract data from Amazon, one of the world's largest and most popular online retailers, and build a robust NER model for entity classification in e-commerce. The initial process is to construct a valid dataset for Amazon product titles, as there are no publicly available free datasets with labeled entities for the titles; work on a statistical approach to pre-process the data to remove noise from it; implement a state-of-the-art supervised classification model using a combination of BiLSTM (Bidirectional Long Short Term Memory), CNN (Convolutional Neural Networks), and CRF (Conditional Random Fields); differentiate the unbalanced data into target groups of strong and weak entities and implement fine-tuned classifiers separately to improve the performance.

1 Introduction and Motivation

Electronic commerce or e-commerce is a service for consumers and producers to buy and sell goods over the internet. Specifically, retail e-commerce is where the product is sold to the customer directly. In the U.S. in 2021, revenue for retail e-commerce is estimated to be 768 billion U.S. dollars and is forecasted to exceed 1.3 trillion dollars by 2025, according to Statista. Having a success this huge in business and a high usage demand from customers and product sellers, it becomes a vital task to improve the experience for both. One of the significant features of online e-commerce stores is the search engine to make the products easily visible to the customers and find the accurate product quickly. It is the most common and preferred way to find products before applying filters to narrow down the displayed items. So, are these search engines intelligent enough to find the customer's desired product? To provide a seamless experience on the search engine, Entity Extraction, also known as Named Entity Recognition (NER), is a search technique that is used in the product discovery domain. The method identifies the entities from unstructured text data, like brand, color, product type, and attributes, to further help understand the information and put it to good use. In this project, we will initially use a Supervised NER method as baseline and work on e-commerce product titles containing the required information to define a product. Later, we will aim to improve the model by using other Distantly Supervised methods for performance comparison to identify the most suitable approach.

In the e-commerce domain, the value of the goods is primarily represented by the product titles, and we selected this data group to continue our research further. When we started working on this project topic, we could only find very few re-



The Mars Orbiter Mission PRODUCT (MOM ORG), informally known as Mangalyaan GPE , was launched into Earth LOC orbit on 5 November 2013 DATE by the Indian Space Research Organisation ORG (ISRO ORG) and has entered Mars LOC orbit on 24 September 2014 DATE . India GPE thus became the first ORDINAL country to enter Mars LOC orbit on its first ORDINAL attempt. It was completed at a record low cost of $74 million MONEY .

Figure 1: NER Sample - Source: link

search works on the e-commerce domain for NER. The latest and best methods to build the NER model are tailored to work for specific fields. This also motivated us to incorporate those methods into the e-commerce space and construct a performance measure.

NER is an information extraction technique to identify and classify named entities in text. It is closely associated with Machine Learning. NER model is a two-step process detecting a named entity and then categorizing the entity. In this project, we will focus on the second step of classifying the entities, which involves models that use statistical methods internally. The list of methods used in NER we came across researching were BiLSTM (Bidirectional Long Short Term Memory) is a type of neural network, CRF (Conditional Random Fields) is a statistical modeling method, PU (Positive Unlabeled) algorithm involves statistics to calculate cost estimations and loss functions, RoBERTa (Robustly Optimized BERT Approach) is an upgradation to BERT (Bidirectional Encoder Representations from Transformers) uses semi-supervised learning, and many other models are tested in this domain that comprises deep knowledge of statistics. The course syllabus and the book Pattern Recognition and Machine Learning by Christoper Bishop contain topics in Classification, Neural Networks, Model Selection, etc., that are beneficial.

2 Problem Description

The search engine is one of the most critical elements of online e-commerce sites since it allows customers to quickly and easily find the right product. It is the most typical and preferred method of finding products before using filters to reduce the number of items presented. Improving user experience is prioritized because E-commerce is a commercial success and has a significant demand for usage. We frequently discover that these search engines lack the intelligence necessary to locate the desired products for customers. Identifying the essential elements in the eCommerce space, such as brand and product type, can assist a search engine find relevant products and provide a delightful buying experience. We discovered very few NER research works on the e-commerce domain inspired us to apply the newest and most effective techniques to the e-commerce industry. Hence we propose a Named Entity Recognition System for E-Commerce Product Titles.

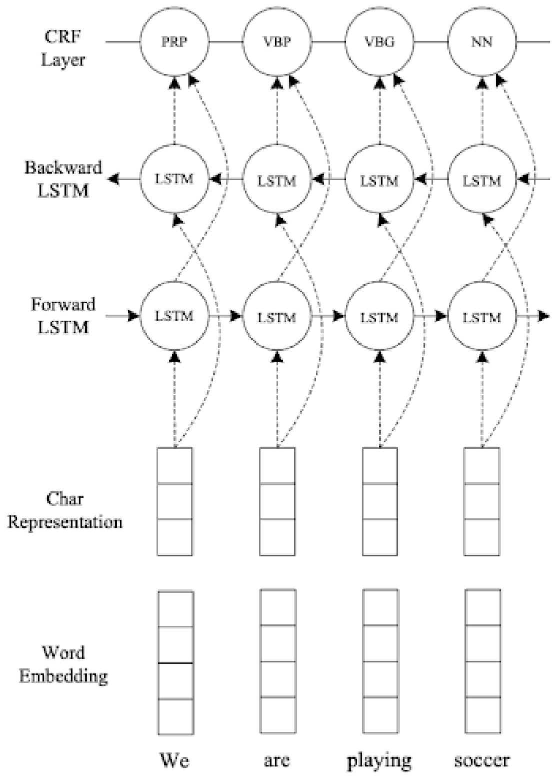


Figure 2: Complete Model - Source: link

3 Methodology

Our model is a combination of BiLSTM, CNN, and CRF layers. Using Character CNN, we extract a character-level representation of a given word for the first layer. The word embedding vector and character-level representation vector are then con-

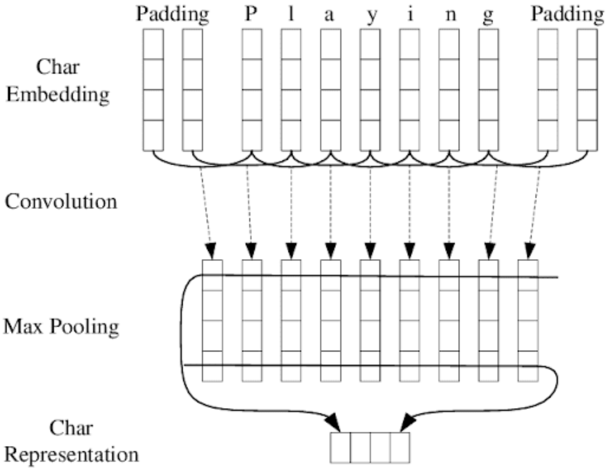


Figure 3: CNN Structure - Source: link

catenated and sent into the BiLSTM network. The CRF layer is then given the BiLSTM output vectors to jointly decode the ideal label sequence.

3.1 CNN Layer

Character-level embedding uses a convolutional neural network to find a numeric representation of words by looking at their character-level compositions. Our motivation behind choosing character-level CNN is composed of three reasons:

- Character-level CNN is suited to noisy content.
- There is no text preprocessing required.
- It handles misspelled or out-of-vocabulary words well.

3.2 BiLSTM Layer

LSTM units are special kinds of RNNs are capable of learning long-term dependencies. Sequence modeling requires a model to persist past information to inform later predictions. Consider an example: “I grew up in France. I speak fluent French.” We need the context of France to be able to predict the name of the language. LSTMs have the ability to remove or add information to the cell state, regulated by three gates:

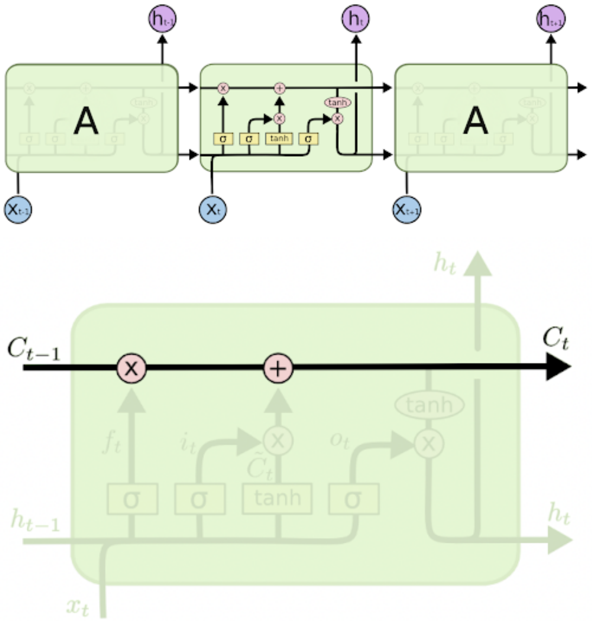


Figure 4: LSTM - Source: link

- Forget gate: throws away the information that is not needed by the model anymore.
- Input gate: stores new information.
- Output gate: sends out the combination of information obtained by the above two gates.

Consider the following two examples:

- “I like apples because they keep me healthy.”
- “I like Apple because of its fancy products.”

How do we know what entity apple is in these cases? We need information from the subsequent words to inform our decision. BiLSTM has two LSTM layers, forward and backward, to predict the output. In the forward layer, activation is from left to right, whereas in the backward layer, activation is from right to left. Prediction is made with the past and future context.

3.3 Vanishing Gradient Problem

Regular RNNs suffer from the problem of vanishing gradients that hampers the learning of long data sequences. The gradients carry information used in the RNN parameter update. When the

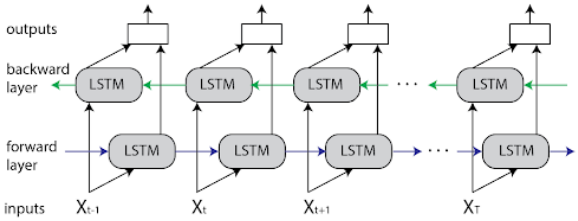


Figure 5: BiLSTM - Source: link

gradient becomes very small, the parameter updates become insignificant. In LSTMs, however, the presence of the forget gate and the additive property of the cell state gradients enables the network to update the parameter so that the different sub-gradients do not vanish.

3.4 CRF Layer

For sequence labeling (or general structured prediction) tasks, it is beneficial to consider the correlations between labels in neighborhoods and jointly decode the best chain of labels for a given input sentence. So we model label sequences jointly using a conditional random field (CRF) instead of decoding each label independently. The probabilistic model for linear CRF defines a family of conditional probability. We employ a linear CRF model in which only interactions between two successive labels are considered. Training and decoding of linear CRFs can be solved efficiently by adopting the Viterbi algorithm.

4 Results

For experimentation, we created our dataset by scraping the Amazon website. The search keyword we used for generating the data is 'Sports team Shirts.' There are 17 entity classes for this search result. For the 17 entity classes, there are 306 matching keywords. For every filter in the entity class, we took roughly 200 web pages and extracted the product titles. As a result, we got around 88,424 product titles. There are some duplicate titles in the titles extracted due to multiple tags assigned to the same title. After filtering and removing the duplicate titles, 32,571 unique product titles are left.

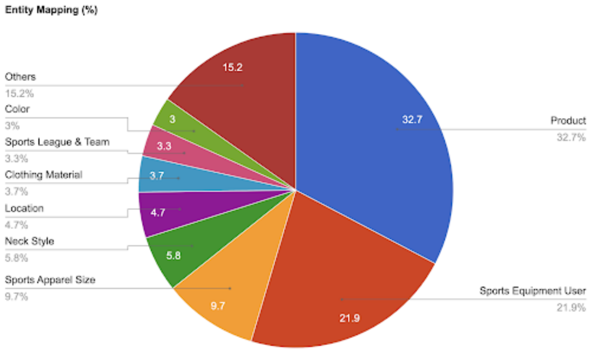


Figure 6: Entity Mapping distribution

Applying String matching algorithms on the titles with the keywords, we extracted the entity mapping frequency of classes as shown in the fig below. We took the top 8 entity classes for our model training and testing, as the other classes' presence in the titles is rare. Among the eight entity classes, we categorized 3 of them as strong and the other five as weak based on the entity mapping. The classification of the strong and weak classes are as follows:

We created three different datasets based on the frequency of the strong entity and weak entity tags. In the first dataset, we considered the frequency of both strong and weak entity class tags and got the top 7500 product titles having a high frequency of both strong and weak entity class tags combined per title. On the other hand, In the second dataset, we have considered only the frequency of strong entity tags and extracted the top 7500 product titles having a high number of tags per title. Only the frequency of weak entity tags in the third dataset is considered. In all three datasets, we have shuffled the extracted titles randomly and selected 5000 titles for training purposes and the remaining 2500 titles for testing. According to the MUC-7 1997 Conference, a 3% mismatch is preferred based on NER human annotator score for Message Understanding. For the 5000 titles that are used for training, 10% of the mismatch or noise is introduced. This assumption is due to the increased manual faulty tagging and decreased quality of tags.

Example title: **"christmas kawaii shirts for women teen girls cute gnome snowflake print crew neck sweatshirts long sleeve comfy blouse"**

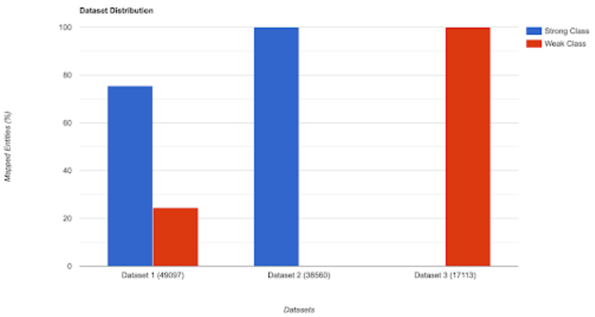
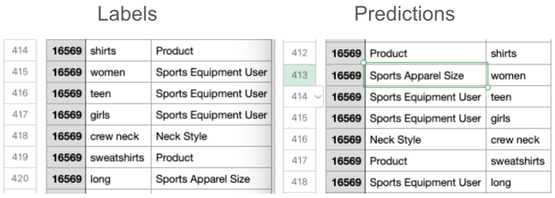


Figure 7: Datasets based on Strong and Weak entity classes



414	16569	shirts	Product
415	16569	women	Sports Equipment User
416	16569	teen	Sports Equipment User
417	16569	girls	Sports Equipment User
418	16569	crew neck	Neck Style
419	16569	sweatshirts	Product
420	16569	long	Sports Apparel Size

412	16569	Product	shirts
413	16569	Sports Apparel Size	women
414	16569	Sports Equipment User	teen
415	16569	Sports Equipment User	girls
416	16569	Neck Style	crew neck
417	16569	Product	sweatshirts
418	16569	Sports Equipment User	long

Figure 8: Sample predictions and labels from our test results

The left figure indicates the training entity mapping and the right table shows the mapping predicted by our model for the example title.

We have considered the average F1 score as an evaluation metric. The table below shows the performance of our model on three different datasets using three different word embedding algorithms.

GloVe.6B.100D word embedding is a pre-computed feature mapping on 6 billion words that are extracted from Wikipedia over 100 meaningful dimensional vector space. The custom-built word embedding is created using the GloVe algorithm but using the tags from product titles in our dataset. It contains 4.4 lakh words and also has 100-dimensional vector space.

Random sampling performed better compared to GloVe in datasets 1 and 2 because the GloVe dataset is built on Wikipedia text and our product titles significantly differ from standard text.

We have fine-tuned our model with different hyperparameters and observed a significant difference with the dropout layer. Since the dataset is small, the dropout layer helped with the overfitting problem and helped in regularization. The F1 score increased significantly for the GloVe dataset up to a dropout value of 0.5, which later decreased slightly.

Table 1: F1 Scores for three datasets on different word embedding vectors

Word Embeddings	Dataset 1	Dataset 2	Dataset 3
GloVe.6B.100d	0.924937	0.905141	**0.923432**
Custom.100d	0.915327	0.904346	0.917639
Random Sampling	**0.928371**	**0.913913**	0.919490

Table 2: F1 Scores for three datasets on different dropout values

Dropout	Dataset 1	Dataset 2	Dataset 3
0	0.924937	0.905141	0.923432
0.2	0.919473	0.903054	0.916325
0.5	**0.979516**	**0.973822**	**0.967744**

5 Related Work

We have aggregated a few technical research papers while working on the topic, namely,

- Distantly Supervised Named Entity Recognition using Positive-Unlabeled Learning - Semi supervised model for NER targeting Unlabelled data. [Pen+19]

- Named Entity Recognition for E-Commerce Search Queries - Compared BiLSTM and Bi-GRU model on customized dataset separated into three forms 2020. [Bha+20]

- Recent Trends in Named Entity Recognition (NER) - Has an aggregated list of research works and model performances in NER region. [Roy21]

- Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training - Experimented with supervised RoBERTa model and other semi-supervised models with Noise-Robust learning 2021. Showed similar results with BiLSTM model. [Men+21]

- End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF - Original paper that we followed gave the best results on CoNLL dataset 2016. [MH16]

- Partially-Typed NER Datasets Integration: Connecting Practice to Theory - Research was focussed on less number of entity mappings per test record could also achive similar performance compared with full presence of entity mappings for text records 2020. [Zhi+20]

- Adaptive Name Entity Recognition under Highly Unbalanced Data - In this paper the dataset was separated based on entity classes Strong and Weak. Different model was chosen for Weak entities and merged with BiLSTM model for Strong entity class to get better results. Our model performance didn't show any significant difference for Strong and Weak Entities separately. So, didn't follow this approach. [NNR20]

- GloVe: Global Vectors for Word Representation - Research paper for GloVe. [PSM14]

6 Conclusion

We created a dataset consisting of Amazon product titles with labeled entities. Generate custom word-embedding vectors using GloVe open-source code used in our evaluations. Evaluated a state-of-the-art NER model and saw the effect of the dropout layer on our testing results. We understood using statistical layers like Convolutional Neural Networks, Bi-Directional Long Short Term Memory, and Conditional Random Fields. Potential future work is improving entities' recognition through unsupervised learning models. As proposed in our plan, we successfully evaluated a supervised NER model and completed the baseline objectives. Our choice of building the dataset from the initial stage was to target a specific area with a heavy requirement for NER improvements. Due to the unavailability of valid data, little research is done, and we could evaluate state-of-the-art NER models in this region. The time our team has spent creating the prototype dataset could be used to evaluate more models and focus on the issues that revolve around the improvements.

References

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: `http://www.aclweb.org/anthology/D14-1162`.

[MH16] Xuezhe Ma and Eduard H. Hovy. "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF". In: *CoRR* abs/1603.01354 (2016). arXiv: `1603.01354`. URL: `http://arxiv.org/abs/1603.01354`.

[Pen+19] Minlong Peng et al. "Distantly Supervised Named Entity Recognition using Positive-Unlabeled Learning". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2409–2419. DOI: `10.18653/v1/P19-1231`. URL: `https://aclanthology.org/P19-1231`.

[Bha+20] Bhushan Ramesh Bhange et al. "Named Entity Recognition for E-Commerce Search Queries". In: 2020.

[NNR20] Thong Nguyen, Duy Nguyen, and Pramod Rao. "Adaptive Name Entity Recognition under Highly Unbalanced Data". In: *CoRR* abs/2003.10296 (2020). arXiv: `2003.10296`. URL: `https://arxiv.org/abs/2003.10296`.

[Zhi+20] Shi Zhi et al. "Partially-Typed NER Datasets Integration: Connecting Practice to Theory". In: *CoRR* abs/2005.00502 (2020). arXiv: `2005.00502`. URL: `https://arxiv.org/abs/2005.00502`.

[Men+21] Yu Meng et al. "Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training". In: *CoRR* abs/2109.05003 (2021). arXiv: `2109.05003`. URL: `https://arxiv.org/abs/2109.05003`.

[Roy21] Arya Roy. "Recent Trends in Named Entity Recognition (NER)". In: *CoRR* abs/2101.11420 (2021). arXiv: `2101.11420`. URL: `https://arxiv.org/abs/2101.11420`.

7 Additional References

1. https://unbxd.com/blog/leverage-entity-extraction-make-ecommerce-search-engine-intelligent/
2. https://www.statista.com/statistics/272391/us-retail-e-commerce-sales-forecast/
3. https://www.expert.ai/blog/entity-extraction-work/
4. https://towardsdatascience.com/named-entity-recognition-ner-meeting-industrys-requirement-by-applying-state-of-the-art-deep-698d2b3b4ede
5. https://medium.com/mysuperai/what-is-named-entity-recognition-ner-and-how-can-i-use-it-2b68cf6f545d
6. https://towardsdatascience.com/named-entity-recognition-applications-and-use-cases-acdbf57d595e
7. https://www.turing.com/kb/guide-on-word-embeddings-in-nlp
8. http://ml.cau.ac.kr/activities/outputs/210803-BIOES
9. https://nlp.stanford.edu/projects/glove/
10. https://towardsdatascience.com/batch-mini-batch-stochastic-gradient-descent-7a62ecba642a
11. https://theaisummer.com/regularization/
12. https://deepai.org/machine-learning-glossary-and-terms/gradient-clipping: :text=Gradient
13. https://d2l.ai/chapter_recurrent-modern/lstm.html
14. https://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_english_score_report.html

8 Code References

1. https://github.com/jayavardhanr/End-to-end-Sequence-Labeling-via-Bi-directional-LSTM-CNNs-CRF-Tutorial
2. https://github.com/ZhixiuYe/NER-pytorch
3. https://github.com/glample/tagger
4. https://github.com/stanfordnlp/GloVe