Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 11, November-2023

Unveiling Movie Titles: Leveraging BERT for Dialogue-Based Extraction

Zehra Sikira
International Burch University, Sarajevo
Bosnia and Herzegovina, BSc

Abstract - **Named entity recognition is the most important phase of any Natural Language Processing task. It has a large application space in information retrieval, document organization, and dialogue systems. To improve an already existing model, created to test the performance of the ReDial dataset, we developed the named entity recognition module for movie title extraction. We used the BERT pre-trained model to capture a small dialogue context, and to use the pre-trained knowledge it has, in an attempt to annotate this data set. Major contributions of this paper are token annotation for the mentioned dataset and the development of a named entity recognition module, which showed good performance on the previously unseen dataset with an f1-score of 98%.**

Keywords - **conversational recommender system, dialogues, movie titles, named entity recognition (key words)**

I. INTRODUCTION

One of the major challenges the modern world is dealing with is information overload. Research of whatever domain becomes impossible. A solution could be provided with Conversational Recommender Systems (CRS) - systems with the goal of providing personalized recommendations with active feedback.

A perfect domain of application of such a system is movies. Nowadays it is uncountable how many movies exist, yet not all of them are popular. With the static Recommender System (RS), there were risks of cold start and popularity bias. Such risks can be minimized with CRS.

For such a system information extraction is a crucial step. Especially in the case of the Recommendation Dialogue (ReDial) dataset intended for CRS training. In dataset conversations, movie mentions are formatted in "@movieId" form, and the system, published together with the dataset, expects user utterances with movies mentioned in such a format which is unnatural [1].

To create an additional module for this system that recognizes and extracts movie titles, we will use Named Entity Recognition (NER) methods. And since the goal is to extract movie titles based on the dialogue context we will fine-tune the pre-trained Transformer-model for this task because they can capture and store contextual information. Additionally, fine-tuning does not require large computing resources as training Large Language Models (LLM) from scratch, therefore the negative environmental impact will be minimized.

II. LITERATURE REVIEW

NER has a wide spectrum of applications. Initially, it was used for information extraction and retrieval, document organization, book indexing, etc. where certain terms and keyphrases would be extracted from documents, which would lead to faster querying [2].

The first NER systems were rule-based, where it was programmed for example that person names start with capital letters and so on. In the scope of data-driven NER approaches, commonly used algorithms were decision trees, hidden Markov models, and hybrid approaches [2].

Based on the problem domain, NER tasks can be open-domain or domain-specific. In the case of open-domain, it analyzes textual data from different domains in an attempt to classify tokens into one of the generic entity types (such as person, location, event, etc.) [3–5]. Such systems have difficulties when classifying domain-specific entities because usually they are used with unlabeled data with the goal of labeling. In [4] for such a task they used clustering techniques. To overcome this difficulty in [3] they used BERT since it has pre-trained knowledge, but it is hard to interpret the model and requires large amounts of data to avoid biases.

Domain-specific NER models initially used external knowledge from different databases or ontologies to learn domain-specific knowledge. But in the 2010s, neural networks became a common algorithm in domain-specific NER [6]. In [7] they used domain-specific dictionaries to label initial data and then used it in combination with Fuzzy-LSTM-CRF to handle tokens with multiple possible labels. Additionally, they used the dictionary in combination with their AutoNER model to handle entity recognition of multi-token entities.

To sum up, this, at first sight, simple task, is of great importance in different types of language processing applications, and depending on the specific use case and domain may become complex.

II. METHODOLOGY

Bidirectional Encoder Representations from Transformers (BERT) became one of the essential models for natural language processing due to its ability to get contextual information and to keep trained knowledge. Both of these are crucial for our system since movie titles consist of everyday words and may contain multiple words. Thus our goal to extract movie titles from dialogues can be accomplished by fine-tuning one of the pre-trained BERT models since the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 11, November-2023

TABLE I. DATASET SIZE

	train	*test*
#conversations	10006	1342
#tokens	1627653	201900

contextual information provided in one dialogue usually is not large.

To do so, we will first have to prepare the dataset for the model, then use preprocessed data for fine-tuning, and finally, we will evaluate our model through analysis of the generated classification report.

A. Dataset Preprocessing

For this task we will use the ReDial dataset. As stated in Table I, the dataset contains 10006 conversations in train and 1342 conversations in the test dataset.

In the preprocessing stage, we first have to replace each movie mentioned in the format "@movieID" with a corresponding movie title (Fig. 1). After that we will concatenate all messages for each conversation, and tokenize the resulting text. Since the "bert-base-uncased" model uses subword tokenization, we will also tokenize movie titles so we can label conversation tokens. In the labeling stage, we will mark conversation tokens that represent movie titles with 1, and the rest of them with 0.

Since for this task, it is not important if the utterance is written by the seeker or recommender, we will concatenate all messages of each conversation into text. After which we tokenized those texts and movie titles.

The next step in preprocessing is to encode conversations' tokens and to convert encodings and labels into tensors. Finally, we will divide the training dataset into training and validation datasets. Validation data will be used for hyperparameters tuning during the training process while the final evaluation will be conducted on previously unseen data from the test set.

For optimal performance, we will use the Hugging Face (HF) dataset structure.

BERT Fine-tuning

We will use the pre-trained BERT model "bert-base-uncased", and fine-tune it for the sequence classification task. In this case, we are talking about binary classification since the token is either part of the movie title or not. BERT base models contain 12 transformer layers and have the dimensionality of hidden representations 768, with 12 attention heads, which results in 110M total parameters. Since we will not train the model from scratch, but just fine-tune it, we will just adjust the model head for our task. We will use an Adam optimizer with a learning rate of 0.00005, and batch size of eight. The model will be fine-tuned over three epochs and with a linear scheduler.

```
Classification Report:
                precision    recall  f1-score   support

Non-Movie Title       1.00      1.00      1.00    178837
    Movie Title       0.98      0.98      0.98     25747

       accuracy                           1.00    204584
      macro avg       0.99      0.99      0.99    204584
   weighted avg       1.00      1.00      1.00    204584
```

Classification Report

Evaluation

To evaluate the performance of the trained model, we will use the test part of the dataset, unseen by the model during the training phase. Then we will generate classification reports with common classification metrics - precision, recall, f1-score, and accuracy.

III. RESULTS AND DISCUSSION

As we can see in Fig.2, our model has an outstanding performance. Firstly, precision, recall, and f1-score are 1.00 for the class of Non-Movie Titles, which implies that each token that is not part of the movie title is correctly predicted. The majority of tokens in each split of the dataset are not part of movie titles, which actually, accurately represent real-life conversations. For another class, a Movie Title, we have lower values for these metrics, which implies that there are tokens that are part of a movie title, however, the model predicted the Non-Movie Title class for them. Even though the f1 score is lower than for the other class, model performance is still outstanding. These results may also imply overfitting.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 11, November-2023

IV. CONCLUSION

Contributions of this paper are token annotation for the ReDial dataset and development of a named entity recognition module that extracts movie titles from dialogues. that showed good performance on a previously unseen dataset with an f1-score of 98%.

Since these results may imply possible overfitting, for future experiments we will enrich this dataset with other data (such as movie reviews).

REFERENCES

[1] LI, Raymond, et al. Towards deep conversational recommendations. Advances in neural information processing systems, 2018, 31.

[2] Borthwick AE. A maximum entropy approach to named entity recognition. search.proquest.com. 1999. Available: https://search.proquest.com/openview/744c41d4

[3] Liang C, Yu Y, Jiang H, Er S, Wang R, Zhao T, et al. BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery; 2020. pp. 1054–1064.

[4] Evans RJ, Street S. A framework for named entity recognition in the open domain. RANLP. 2003. Available: https://www.torrossa.com/gs/resourceProxy?an=5015997&publisher=FZ4850#page=280

[5] Bowden KK, Wu J, Oraby S, Misra A, Walker M. SlugNERDS: A Named Entity Recognition Tool for Open Domain Dialogue Systems. arXiv [cs.CL]. 2018. Available: http://arxiv.org/abs/1805.03784

[6] Yadav V, Bethard S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. arXiv [cs.CL]. 2019. Available: http://arxiv.org/abs/1910.11470

[7] Shang J, Liu L, Ren X, Gu X, Ren T, Han J. Learning Named Entity Tagger using Domain-Specific Dictionary. arXiv [cs.CL]. 2018. Available: http://arxiv.org/abs/1809.0359

