Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 198 (2022) 373–378

Procedia Computer Science

www.elsevier.com/locate/procedia

The 2nd International Workshop on Artificial Intelligence for Natural Language Processing (IA&NLP'21)
November 1-4, 2021, Leuven, Belgium

Named Entity Recognition applied on Moroccan tourism corpus

Ibrahim Bouabdallaoui*, Fatima Guerouate, Samya Bouhaddour, Chaimae Saadi, Mohamed Sbihi

Laboratory of System Analysis, Information Processing and Industrial Management - EST Salé
Mohammed V University in Rabat, Avenue le Prince Héritier, Salé 11060, Morocco

Abstract

The tourism industry has particularly evolved in the era of social media, and due to the huge amount of data available on the web, tourism forums have become spaces for exchanging information about many topics concerning a country's tourism and hospitality, which has helped to establish a bank of knowledge that might encourage artificial intelligence researchers to follow some research tracks that could lead to interesting insights for decision makers in the sphere of tourism. This paper is an application of artificial intelligence within the tourism industry in the context of Morocco, whereby we carried out natural language processing transformers to classify entities (i.e., Named Entity Recognition) in text that was collected from the Moroccan forum in TripAdvisor.

Keywords: Named Entity Recognition; transformers; Moroccan tourism; web scraping

1. Introduction

Web forums are powerful online message boards whereby users are confident and enjoy to hold discussions via message post on relevant threads, and to share useful information that resolves common problems tourists may face, and furthermore to discuss their opinions freely. Tourism forums have existed since the idea of creating forums was conceived and implemented, the intention was that the forum users discuss only positive tourism plans and experiences, and for instance good accommodation and touristic areas to travel to [1]. TripAdvisor is one of the most

* Corresponding author. Tel.: +212-663-110-508
E-mail address: bd.ibrahim@hotmail.com

10.1016/j.procs.2021.12.256

popular Travel Social Network websites, it reached 884 million reviews/opinions in 2020 (as the last TripAdvisor statistics performed by Statista Research Department), which therefore makes TripAdvisor the most enriched website in terms of both data and trustworthiness. Tourists are able to plan their trip, check information, take heed of the experiences of others, and establish ranking lists.

The latest technologies involved in Natural Language Processing (NLP) have inspired and indeed ensured significant attention due to their efficiency regarding language modeling. The power of neural networks has achieved interesting results in terms of classification, generation and clustering of text. The core of these algorithms relies upon mathematical approaches that make understanding text patterns more accurate. However, context vectors have proven to be a "bottleneck" for many types of simple models, which make text mining models building quite difficult regarding long sentences. Hopefully, attention models arise to solve such issues [2], since it has improved the quality of machine translation systems. Attention allows the model to focus on relevant parts of the input sequence as necessary.

Named Entity Recognition (NER) is an information retrieval subtask that seeks to locate and classify named entity mentioned in unstructured text [3] into predefined categories, locations, medical codes, expressions of time, quantities, monetary values, percentages, etc...

2. Related work

A work had been published to automate the labeling of tourism data [4], and particularly hotels data from TripAdvisor and Hotels.com concerning 8 provinces in Thailand. A vocabulary is built for handling multiword using the multiword tokenizer to prepare the annotation training for recognizing new entities and keywords, such as locations and facilities. Two extracting methods had been used in this paper: the first one being training sentences, and finding sentences that contain these named entities, and then prepare results for text summarization, the second one is the relation type extraction when there are two another approaches for recognizing relations, the first approach is to use dependency parsing, and the second one is to train keywords as new tags. To recognize both named entities and relation names at the same time, indication words has been created, which imply the location, and facility in the sentences respectively, and then a sentence tokenization is performed. To then utilize this pre-processing pipeline in BERT [5] NER model, the BIO-tagging (which is a common markup format for marking up tokens in a segmentation task) can then be performed: this is known as Computational Linguistics.

Another work has been inducted using NER based on semi-supervised learning with the purpose of identifying new tourism destinations [6], the main entities that the research aims to predict are "Nature", "Place", "City", "Region" and a negative class which refers to an unknown entity. Processing data in this paper has comprised 4 steps: the first one being the sentence splitter to split sentence to chunks, the second one being tokenization regarding words and punctuation, the third one is the Part-of-Speech that shows whether the word is a noun, adjective, verb, etc…, the last step is the 'NP chunker' which is used to select candidates based on Part-of-Speech formed entities. The result of this data preparation step is stored in an entity candidate list, and then fed to the Naïve Bayes Classifier to detect whether the token is an area entity or not, and furthermore the YATSI SSL algorithm does the work of classifying the Named Entity.

3. Methodology

This section details the methodology we have implemented in order to build 3 NLP models based on transformers [7]: BERT, RoBERTa [8] and XLM-RoBERTa [9], applied to the Moroccan forum in TripAdvisor, which we have collected using web scraping techniques. The main aim of this work, is to extract important Moroccan tourism entities, and to understand the meaning of sentences following the positioning of relevant targeted words. As a result, these models would help us to extract vital information from complex and non-labelled text, which poses a real challenge in the realm of Natural Language Processing.

3.1. Dataset

We have focused on TripAdvisor forums, due to TripAdvisor's reliability and enriched content [10] and the huge community who are interested in tourism in Morocco. The Moroccan forum contains 46120 topic threads, whereby each thread has one message or more, some topic threads have more than one hundred messages, these messages relate to: recommendations, optimal plans, notices or complaints on hotels, restaurants, transports, services…

Using BeautifulSoup package in Python 3, we managed to loop over all the pages of the forum, and loop over all threads on a page; each thread has a title, an author and location, a timestamp and its own content. The data collected is raw text, which is unstructured data and since it is useless to extract it and put it in a text file, therefore a semi-structured schema is proposed to organize this data, and store it in a formal way using JSON file format, in order to have an easy access while loading data and perform exploration and pre-processing pipelines.

On the other hand, restaurant data has been collected in each city of Morocco, each restaurant is named, as is its address and geographical position. The names of shopping facilities have been collected too with their types from TripAdvisor website. Also, we took 3 databases available on the Moroccan government open data sources, the first database contains all of the regions of Morocco (ranging from big cities to small villages), and taking into consideration abbreviations of cities (e.g., Casablanca and Casa), the second one contains details about all of the hotels in Morocco such as names, caliber (i.e., class/star rating), address and phone; the third database contains names of the touristic guides themselves, and indeed their specialty and where they are located. Then we created a dictionary containing local Moroccan keywords, such as gastronomy, religion places, outfits and transport options.

The logic behind including other sources of data, is to have knowledge of many keywords in the forum, and to have logged full notation, which could help models to compute the meaning of sentences more easily without returning unknown outputs.

3.2. Data pre-processing

After collecting useful data, we then utilize it as input for our models. This is a compulsory step in order to prepare data, indeed this step is primordial and it is imperative to action it before doing any analysis task on text. Since we have a json file containing semi-structured data, we then loaded data into a dictionary data structure, to have an easy processing through regular python basics such as looping and conditions, and perform fast runtime execution.

In order to pre-processing data, we performed as a first step, the conversion of all target words into lower case text, therefore making the text normalized, and then eliminate handles and URLs, and punctuations. Spacy package is loaded to tokenize strings into words, however before executing this step, we created an entity ruler, that contains all the keywords that we had collected from the databases (i.e., regions, hotels, restaurants, shopping places, gastronomy, religion, outfits, transports and touristic guides), and consider it as sentence patterns, then we tokenized strings into words, using an existing spacy model, which contains all of the patterns of the English language. The outputs were then stored in JSON format, whereby each topic thread has its own JSON file, that contains a vector of tokens, a vector of tags, vector of label of the text, and the entity vector.

In this paper, we are interested in tokens and tags which are the Named Entity Recognition for information gathering purposes, that we will utilize in the various models, and we will discuss these further in the following subsections.

3.3. BERT model:

One of the latest milestones in the development of NLP, is the release of BERT (Bidirectional Encoder Representations from Transformers), an event described as marking the start of a new era for NLP. BERT is a model that has broken several records for language-based modeling tasks. BERT builds on a number of ideas that have recently emerged in the NLP community, including semi-supervised sequential learning [11], ELMo [12], ULMFiT [13], the OpenAI transformer [14] and the Transformer.

Like the Transformers encoder [15], BERT takes a sequence of words as input that goes up the stack. Each layer applies self-attention and transmits its results to a feed-forward network, then forwards them to the next encoder. Finding the right way to train a stack of encoders is a complex obstacle that BERT solves by adopting a concept of a "Masked LM". This procedure consists of randomly taking 15% of the input tokens and then hiding 80% of them, replacing 10% with another completely random token (another word) and doing nothing in the case of the remaining 10%. The goal is that the model correctly predicts the original modified token (via cross-entropy loss). The model is therefore forced to maintain a distributional contextual representation of each input token [16].

In term of output, each output provides a vector of hidden size (768 in BERT Base). This vector can now be used as an input for a classifier of our choice.

In this paper, we are interested in fine-tuning tokens as tags using the Named Entity Recognition technique.

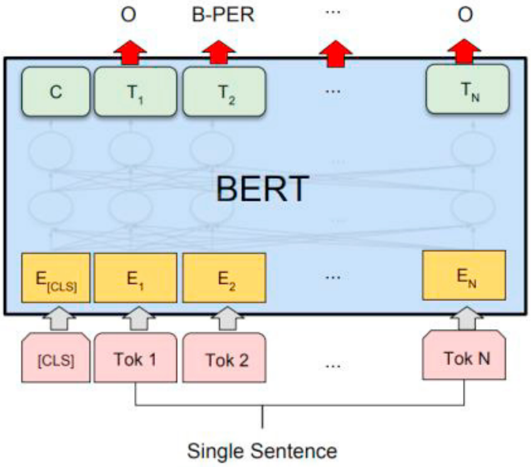

Fig. 1. BERT architecture for token classification

3.4. RoBERTa model:

RoBERTa is a model built on BERT language masking strategy, makes a robustly optimized method for pretraining natural language processing systems. The system learns to predict hidden sections of text intentionally within otherwise non-annotated language examples. RoBERTa modifies key hyperparameters in BERT, including removing BERTs next-sentence pretraining objective, and training with much larger mini-batches and learning rates. This new feature improves the masked language modeling objective in comparison to BERT and leads to better downstream task performance. RoBERTa performs good results in term of magnitude with more data that requires a longer amount of time of execution.

Results show that tuning the BERT training procedure can significantly improve its performance on a variety of NLP tasks while it also indicated that this overall approach remains competitive versus alternative approaches. This work shows the potential of self-supervised training techniques to match or exceed the performance of more traditional, supervised approaches.

3.5. XLM-RoBERTa model:

XLM-RoBERTa belongs to RoBERTa family, it is trained on a multi-lingual language modeling objective [17] using only monolingual data, which entails those streams of text samples are taken from all the languages, and the model performs a training task to predict masked tokens in the input. It is released in November 2019 by Facebook AI as an update to the original XLM-100 model [18]. These two models are transformer-based language, and they rely upon the Masked Language Model [19] objective and they are capable of processing text from 100 separate languages. There is no Next Sentence Prediction as BERT or Sentence Order Prediction like ALBERT [20]. XLM-Roberta uses the one large shared Sentence Piece model [21] to tokenize instead of having a slew of language specific tokenizers as was the case in XLM-100.

4. Results and Discussions

After preparing the dataset, and identifying tags as label indexes (which entails that we attributed values for all the tags that we proposed in the last subsection) and grouped sentences in accordance with their tokens and tags, then we split the data to a list of sentences and a list of labels in order to be ready for the suitable model Tokenizer, because tokenizing words using a wrong model tokenizer that it is not compatible with the model, would definitely cause errors.

It is a compulsory task to go through the process above for every attention model preparation, is conversion of tokens to indexes, building attention masks on the basis of these indexes, and finally assign tags using the labels list. Another task that must be done, is to split data to 99% for the train set and 1% for the validation set so that the model can learn the maximum possible features that are available on the text, especially because we have text that contains many topics and many features that the model must learn. After converting both divided datasets to tensors, and instantiate the suitable model for training, using 8 epochs, AdamW optimizer for fine tuning the transformer, a variant weight decay to find the optimal point and a scheduler to execute these tasks. The function of activation of the 3 models is SoftMax. Due to the variety of tags that we had; this type of function would return the suitable prediction for classification after applying the argmax to find the arguments that gives us the maximum value.

For each model, loss values of validation set are stored, and plotted as follows:

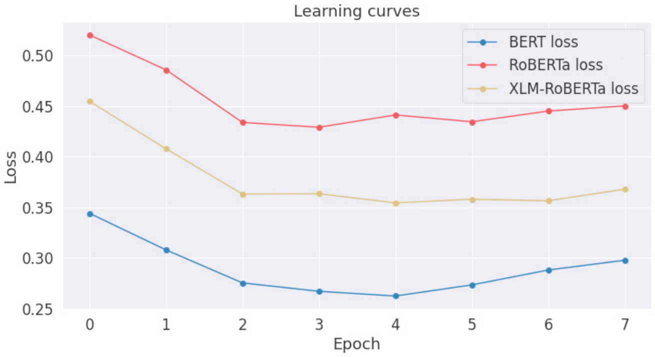

Fig. 2. Validation loss curves

As we can see, BERT [5] model, RoBERTa [8] model and XLM-RoBERTa [9] have respectively a validation loss that is initialized on 0.55, 0.46, 0.34 and drains to the optimal point in the 5th iteration, and then it increases again slightly, that means that the model might go for an overfitting if the number of epochs is above 6.

The following table shows results of the best point of the plot, using a model checkpoint that saves the model on its top performance. Each model is evaluated using accuracy, precision, recall and F1-Score metrics:

Table 1. Performance of models

Metrics	BERT	RoBERTa	XLM-RoBERTa
Accuracy	0.877994	0.884426	**0.900761**
Precision	**0.705826**	0.699895	0.686763
Recall	**0.730067**	0.713686	0.707940
F1-Score	**0.705015**	0.694455	0.685398

In the table above, the BERT [5] model gave highest score in F1-Score in comparison with the RoBERTa [8] family models, which means that BERT [5] is predicting values with low false positive and low false negative, since the Recall and Precision are the highest values, however, regarding the accuracy it reaches the lowest value. On the other hand, it is more likely close to the other metrics, which means that the result of BERT [5] is more significant than the others. For the XLM-RoBERTa [9] accuracy, it proves that this model can perform best prediction in only true positives and true negatives, and its much further higher than the F1-Score value.

5. Conclusion and future work

In this paper, three approaches of transformers were utilized: BERT [5], RoBERTa [8], XLM-RoBERTa [9], and have been applied to the tourism dataset in order to perform token classification for Named Entity Recognition, we have considered the results of the previous sections and we have noticed that BERT [5] performs accurate and robust predictions. The result of this work, could be used for information retrieval, to understand the context of a sentence that refers to Moroccan tourism, and also to create a large database of useful Moroccan entities that the tourists aspire to, during their trip to Morocco. Our next work will be a combination of several transfer learning models, and Recurrent Neural Networks, with the purpose of performing Named Entity Recognition and Questions-Answers models.

6. References

[1] Leung, Daniel, Law, Rob, van Hoof, Hubert, and Buhalis, Dimitrios. (2013) "Social Media in Tourism and Hospitality: A Literature Review" *Journal of Travel & Tourism Marketing* **30**: 3–22.

[2] Ashish, Vaswani, Noam, Shazeer, Niki, Parmar, Jakob, Uszkoreit, Llion, Jones, Aidan N., Gomez, Lukasz, Kaiser, and Illia, Polosukhin. (2017) "Attention is all you need". *Advances in neural information processing systems*: 5998–6008.

[3] Li, Jing, Sun, Aixin, Jianglei, Han, and Li, Chenglian. (2020) "A Survey on Deep Learning for Named Entity Recognition" *IEEE Transactions on Knowledge and Data Engineering*: 1-1.

[4] Chantrapornchai, Chantana, and Tunsakul, Aphisit. (2019) "Information extraction based on named entity for tourism corpus" *16th International Joint Conference on Computer Science and Software Engineering (JCSSE). IEEE*: 187-192.

[5] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. (2018) "Bert: Pre-training of deep bidirectional transformers for language understanding" *arXiv preprint arXiv:1810.04805*.

[6] Saputro, Khurniawan Eko, Sri Suning Kusumawardani, and Silmi Fauziati. (2016) "Development of semi-supervised named entity recognition to discover new tourism places" *2016 2nd International Conference on Science and Technology-Computer (ICST). IEEE*: 124-128.

[7] Kitaev, Nikita, Kaiser, Łukasz, and Levskaya, Anselm. (2020) "Reformer: The efficient transformer". *arXiv preprint arXiv:2001.04451*.

[8] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin (2019). "Roberta: A robustly optimized bert pretraining approach". *arXiv preprint arXiv:1907.11692*.

[9] Conneau, Alexis, Khandelwal, Kartikay, Goyal, Naman, Chaudhary, Vishrav, Wenzek, Guillaume, Guzmán, Francisco, Grave, Edouard, Ott, Myle, Zettlemoyer, Luke, and Stoyanov, Veselin. (2019) "Unsupervised cross-lingual representation learning at scale". *arXiv preprint arXiv:1911.02116*.

[10] Ma, Shihan et Kirilenko, Andrei. (2021) "How Reliable Is Social Media Data? Validation of TripAdvisor Tourism Visitations Using Independent Data Sources". *Information and Communication Technologies in Tourism 2021*. Springer, Cham: 286-293.

[11] Dai, Andrew M., and Le, Quoc V. (2015) "Semi-supervised sequence learning". *Advances in neural information processing systems* **28**: 3079-3087.

[12] Peters, Matthew E., Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, and Zettlemoyer, Luke. (2018) "Deep contextualized word representations". *arXiv preprint arXiv:1802.05365*.

[13] Howard, Jeremy, and Ruder, Sebastian. (2018) "Universal language model fine-tuning for text classification". *arXiv preprint arXiv:1801.06146*.

[14] Radford, Alec, Narasimhan, Karthik, Salimans, Tim, and Sutskever, Ilya. (2018) "Improving language understanding by generative pre-training".

[15] Raganato, Alessandro, Tiedemann, Jörg. (2018) "An analysis of encoder representations in transformer-based machine translation". *Proceedings of the 2018 EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP*. The Association for Computational Linguistics.

[16] Levy, Omer, Remus, Steffen, Biemann, Chris, and Dagan, Ido. (2015) "Do supervised distributional methods really learn lexical inference relations?". *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*: 970-976.

[17] Cho, Jaejin, Baskar, Murali Karthick, Li, Ruizhi, Wiesner, Matthew, Sri Harish, Mallidi, Yalta, Nelson, Karafiát, Martin, Watanabe, Shinji, and Hori, Takaaki. (2018) "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling". *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE: 521-527.

[18] Conneau, Alexis, and Lample, Guillaume. (2019) "Cross-lingual language model pretraining". *Advances in Neural Information Processing Systems*, **32**: 7059-7069.

[19] Song, Kaitao, Tan, Xu, Qin, Tao, Lu, Jianfeng, and Liu, Tie-Yan (2019) "Mass: Masked sequence to sequence pre-training for language generation". *arXiv preprint arXiv:1905.02450*.

[20] Lan, Zhenzhong, Chen, Mingda, Goodman, Sebastian, Gimpel, Kevin, Sharma, Piyush, and Soricut, Radu. (2019) "A Lite BERT for Self-supervised Learning of Language Representations". *arXiv preprint arXiv:1909.11942*

[21] Kudo, Taku, and Richardson, John. (2018) "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing". *arXiv preprint arXiv:1808.06226*.