Project Details

**Objective**

The objective of the MOKB(Moving Objects Knowledge Base)-construction project is to extract structured information from unstructured text written in natural language. The purpose of doing so is to store the structured information in a database and query it using database access languages such as SQL, which cannot be applied to free text. For example, consider an email chain in which employees discusses a planned trip. The MOKB-construction software will extract from the text a record with fields:

Who (the entity taking the trip, maybe a group of people), From-location, Departure-time, Departure-date, To-location, Arrival-time, arrival-date etc.

By extracting such information from the emails of a company a relation of trips can be extracted, and later queried by SQL. Observe that the email texts cannot be queried in this fashion.

For finding the travel records, we are planning to use semi-supervised learning which is an approach to machine learning which combines a small amount of labelled data with unlabelled data during training.

For this, we need to label some emails according to our requirements for a trip record.  Labelling means scanning through the document and naming the entities, a process termed as Named Entity Recognition(NER).  We can use off the shelf NER tools like Spacy, NLTK, OpenNLP etc or can manually name the entities according to our requirements. In our case labeling means identifying the trip details such as traveler name, departure/arrival location, date and time. This labelled set is called the seed set(labelled). After obtaining the seed set, we have to train a classifier model using the pretrained language models like BERT(Bidirectional Encoder Representations from Transformers) . Then we predict the labels for rest of the email set using this model.

**Introduction**

Automatic named entity recognition (NER) is one of the primary tasks in natural language processing[1]. Many well-known NER datasets consist of documents with three types of named entities labeled as persons(PER), organizations(ORG), and locations(LOC or GPE). We see a lot of papers[2][3] discussing NER used in bio-medical field. For these types of named entities, the state-of-the-art NER methods usually give impressive results. However, in specific domains, the performance of NER systems can be much lower due to necessity to introduce new types of entities, to establish the principles of their labeling, and to annotate them consistently.

NER can be done in many ways where the state-of-the-art method is through Transformers models. A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence. In recent years, the transformer model has become one of the main highlights of advances in [deep learning](https://bdtechtalks.com/2019/02/15/what-is-deep-learning-neural-networks/) and deep neural networks. It is mainly used for advanced applications in natural language processing. Transformers, introduced in the 2017 paper “[Attention Is All You Need](https://arxiv.org/abs/1706.03762),” made two key contributions. First, they made it possible to process entire sequences in parallel, making it possible to scale the speed and capacity of sequential deep learning models to unprecedented rates. And second, they introduced “attention mechanisms” that made it possible to track the relations between words across very long text sequences in both forward and reverse directions[5][6].

In this project, we are doing NER of the email dataset using transformer models. We use Bidirectional Encoder Representations from Transformers (BERT) which is a pretrained language model which helps with training the model.

**Dataset**

We have the publicly available Podesta email dataset in [4]. It has about 57,000 emails. The email messages must be preprocessed first, then we need to find the travel relevant emails from the set using a keyword-based search analysis (see Figure 1). After getting the travel relevant emails, annotation must be done. Annotation labels are explained in the Table 1.

How do I filter the travel relevant emails?

Read through each email and check if each sentence contains an entity like PERSON and LOCATION (using off the shelf NER tools like Spacy). For sentences that contain these entities, a keyword-based search is done to detect if this email is relevant to our project. The keywords involved are as follows: 'travel','landed','trip','flying','visit','journeyed','en route','tour','pilgrimage','outing','excursion','going','voyage', 'campaign', 'campaigning' etc. If any of these keywords are present in the selected email, then it is added to the relevant set.

Rules of labelling travel-relevant emails

We label trips based on the following rules or a trip is valid if the following rules are satisfied.

1. A trip is labelled only if it is an inter-city travel. We are looking for flight trips.
2. A trip location should always be a city/state/country. It should not be an organization/building/university etc. For example: “Bill Clinton is going to White House.”. Here, you cannot label “White House” as a TOLOC since it is not a city/state/country name.
3. A trip can be of any 4 types:
4. One way trip: A traveler who takes one way trip is labelled as ‘TRAVELER\_ONEWAY
5. Visits: A traveler who takes a visit to a location(city/state) is also labelled as ‘TRAVELER\_ONEWAY. This is similar to one-way trip except that the departure location and departure date/time will be NULL. But it should have TOLOC as a city/state
6. Round trip: one who takes a round trip is labelled as ‘TRAVELER\_ROUNDTRIP’ and
7. Multi city trips.: A traveler who takes trips to multiple locations is labelled as ‘TRAVELER\_MULTI’.

The departure location, arrival location, departure date/time and arrival date/time of one-way trip is labelled as in Table 1. See table 2 and table 3 for the labels for round trips and multiple location trips.

Table 1. Annotation Labels of one-way trip/visits

|  |  |
| --- | --- |
| **Entity type** | **Label** |
| Travelling Person or a Group | TRAVELER\_ONEWAY |
| Departure Location | FROMLOC |
| Departure Date | DEPARTDATE |
| Departure Time | DEPARTIME |
| Arrival Location | TOLOC |
| Arrival Date | ARRIVDATE |
| Arrival Time | ARRIVTIME |

Table 2. Annotation labels of round-trip

|  |  |
| --- | --- |
| **Entity type** | **Label** |
| Travelling Person or a Group | TRAVELER\_ROUNDTRIP |
| Departure Location | FROMLOC |
| Departure Date | DEPARTDATE |
| Departure Time | DEPARTIME |
| Arrival Location | TOLOC |
| Return Date | RETURNDATE |
| Return Time | RETURNTIME |

Table 2. Annotation labels of multi-trips

|  |  |
| --- | --- |
| **Entity type** | **Label** |
| Travelling Person or a Group | TRAVELER\_MULTI |
| Departure Location | FROMLOC |
| Departure Date | DEPARTDATE |
| Departure Time | DEPARTIME |
| Arrival Location (can be more than 1 location) | TOLOC |
| Arrival Date | ARRIVDATE |
| Arrival Time | ARRIVTIME |

For example: In trips with multiple location like this; “Bolton is travelling to Illinois, Arkansas and Florida”

Bolton- TRAVELER\_MULTI

Illinois- TOLOC

Arkansas- TOLOC

Florida- TOLOC

Get the email body using Email Parser

Annotate the valid trips in the emails

Train the NER model using transformers

Combine the email chain documents into a single document

Remove punctuation and hyperlinks

Remove unwanted spaces

Remove URLS, HTML tags

Pre-Processing steps

Get the travel relevant emails from the dataset

Test and evaluate the model on unlabeled emails

Prepare the labelled set into a specified format for training

1

2

3

7

6

5

4

8

Figure 1 Pipeline of the project

**Explanation of the pipeline**

In figure 1, the step 1 combines the emails and responses with the same subject and merges into a single email chain document. So, from 57,000 emails, we have 31,790 email chain documents after combining the emails and responses of the same subject.

In Step 2 and 3, we get the email messages from each email chain document and do some cleaning like removing punctuations, unwanted spaces etc.

In Step 4, from the cleaned email chain documents; we need to label some number of emails for using as a seed set for training the model.

For this, we filter out travel relevant emails from the entire set using keyword-based search. Then we manually read through the travel relevant emails to check for any valid trips.

In Step 5, if we find a valid trip, we label/annotate the attributes like traveler, from location, to location, departure date, departure time, arrival date, arrival time as TRAVELPER, FROMLOC, TOLOC, DEPARTDATE, DEPARTIME, ARRIVDATE, ARRIVTIME respectively.

In Step 6, the labelled seed set is saved into a tokenized format suitable for training where each token is either labelled as one of them in Table 1. Also, it should follow a BIO tagging scheme to label entities. An example is shown in Figure 2.

|  |  |
| --- | --- |
| **Hillary** | **B-TRAVELPER** |
| **Clinton** | **I-TRAVELPER** |
| **is** | **O** |
| **Travelling** | **O** |
| **To** | **O** |
| **Chicago** | **B-TOLOC** |
| **.** | **O** |

Figure 2. Annotation of entities/tokens using BIO tagging scheme (B-Beginning, I-Inside, O-Outside of an entity)

We need to annotate atleast 1000 email chain documents for our project.

In Step 7, after getting the labelled seed set, we can train the model using transformers (using BERT) or any other state-of-the-art methods for named entity recognition.

We may also try another tool[7] for training the model. For this tool, we may require another format of the labelled data as a sequence.

In Step 8, we can test and evaluate the model to predict the labels for rest of the unlabeled email documents.

**Responsibilities of interns**

1. **The main responsibility is labelling the email chain documents and saving into the specified format(tokenized). We are aiming to get at least 1000. The more, the merrier.**
2. In Step 4, I have used a keyword-based approach for filtering the travel relevant emails from the entire email set. Here, you can make improvements by writing a better script for getting travel relevant emails. This helps you get more emails for labelling.
3. In Step 7, I am planning to train the seed set using transformers and BERT. You must do a literature survey on the methods of training the custom named entity recognition model. (BERT+NER+NLP)
4. In Step 7, we can also try a slot filling tool [7] which is used for airline travel and try to use for our domain.

Note: Instructions for installing [7] is [here](https://docs.google.com/document/d/13aWF186I6cBieFVO8SsARfUI-xdoNimSMh6jK643OEs/edit).

My github link: <https://github.com/avijay6/EmailNERTrips>

**References**

1. https://www.analyticsvidhya.com/blog/2021/11/a-beginners-introduction-to-ner-named-entity-recognition/
2. Perera N, Dehmer M, Emmert-Streib F. Named Entity Recognition and Relation Detection for Biomedical Information Extraction. Front Cell Dev Biol. 2020;8:673. Published 2020 Aug 28. doi:10.3389/fcell.2020.00673
3. A. Sniegula, A. Poniszewska-Maranda, L. Chomatek. Study of named entity recognition methods in biomedical field. Proced. Comp. Sci., 160 (2019), pp. 260-265, [10.1016/j.procs.2019.09.466](https://doi.org/10.1016/j.procs.2019.09.466).
4. <https://wikileaks.org/podesta-emails/>
5. V., Sowmya & Balasubramaniam, Ramya. (2022). What do we Really Know about State of the Art NER?.
6. <https://medium.com/geekculture/3-techniques-to-perform-named-entity-recognition-ner-on-text-data-ec1e91e3a8aa>
7. <https://github.com/monologg/JointBERT>