Structured Sentiment Analysis

Group details :

Group name : pentagon

- Chikkula. Lokesh - 18CS10010
- Ch.Sunil Kumar - 18CS30013
- Divyanshu kumar - 18CS30016
- Rishab Maurya - 18CS30037
- K. Dileep - 18CS10030

Task Description :

Task is to extract a quadruple from the given sentence :

- Holder or source
- Target
- Expression
- Polarity or sentiment

We were provided with data in different language like catalan,basque,Spanish,English etc.. .The analysis of baseline models results were described below.

Approach :

Data preprocessing : We split each sentence into word tokens and assign a label to each token. We generated labels following the preprocessing steps from “sequence labeling” baseline model.

Total number of labels = 11

Ex : B-targ-positive ---> This means beginning of target which has positive polarity

Ex : I-exp-negative ---> This means inside of expression which has negative polarity

Example sentence : "text": "I know some people complained .",
"opinions": [{"Source": [["I"], ["0:1"]],
"Target": [["some people"], ["7:18"]],
"Polar_expression": [["complained"], ["19:29"]],
"Polarity": "Negative", "Intensity": "Standard"}]

This is preprocessed to :

"text": ["I", "know", "some", "people", "complained", "."],
"sources": ["B-holder", "O", "O", "O", "O", "O"],
"targets":["O", "O", "B-targ-Negative", "I-targ-Negative", "O", "O"],
"expressions": ["O", "O", "O", "O", "B-exp-Negative", "O"]}

Training : We used `BertTokenizerFast.from_pretrained('bert-base-uncased')` for tokenizing each sentence.Maximum length of the sentence we have encountered in the training set is 127.So, we padded each sentence to get the length of 128.

Model = BertForTokenClassification('bert-base-uncased,num_labels)

Number of layers in model = 11

We have used AdamOptimizer,loss function to be CrossEntropyLoss function

```
TRAIN_BATCH_SIZE = 32

VALID_BATCH_SIZE = 16

LEARNING_RATE = 1e-05

MAX_GRAD_NORM = 10
```

Results :

Number of epochs = 8

Final training Accuracy = `0.8845185168942342`

Final training loss = `0.36533550752533805`

Final validation accuracy = `0.7620506741334987`

Tuple and span generation :

We iterated each sentence in the test.json file. For each word in the sentence we have predicted labels as mentioned above(11 labels like B-exp-positive , B-targ-negative ...) . From the labels we have generated the span of each opinion by combining beginning(B) and Inside(I) of the appropriate predicted labels.There are few labels without beginning tag(B), we have ignored those types of labels. After calculating the span of each opinion, we have generated tuples.If there are multiple positive or negative targets, we are placing them in a single list[not considering it as an extra tuple]. We have done the same for expressions and holders as well.

We have maintained a mega list of dictionaries,where each dictionary corresponds to each sentence with key values of “sent_id”,”text” and “opinions” .The values of opinions is also a list of dictionaries, where each dictionary is a tuple. With key values as “Source”,”targets”,”Polar_expression” , “Polarity” and “Intensity”.we haven’t stressed much on Intensity of the opinion.So, it affects the accuracy. As of now ,we have only worked on “opener_en” dataset i.e english, So,we weren’t able to submit the task on codalab for evaluation.

Example :


```
{"sent_id":
"opener_en/kaf/hotel/english00106_74e62e551e5b441356b001c5f7ca2070-3"
, "text": "While the hotel is in great condition , the service ruined
my experience .", "opinions": [{"Source": [["my experience"],
["59:72"]], "Target": [["the service"], ["40:51"]],
"Polar_expression": [["ruined"], ["52:58"]], "Polarity": "Negative",
"Intensity": "Standard"}, {"Source": [[], []], "Target": [["the
hotel"], ["6:15"]], "Polar_expression": [["in great condition"],
["19:37"]], "Polarity": "Positive", "Intensity": "Standard"}]}
```


Our extracted tuple :


```
{"sent_id":
"opener_en/kaf/hotel/english00106_74e62e551e5b441356b001c5f7ca2070-3"
, "text": "While the hotel is in great condition , the service ruined
my experience .", "opinion": [{"Source": [["my"], ["59:61"]],
"Target": [["the hotel"], ["6:15"]], "Polar_expression": [["in"],
["19:21"]], "Polarity": "Positive", "Intensity": "Standard"},
{"Source": [["my"], ["59:61"]], "Target": [["the hotel"], ["6:15"]],
"Polar_expression": [["great condition"], ["22:37"]], "Polarity":
"Positive", "Intensity": "Standard"}, {"Source": [["my"], ["59:61"]],
"Target": [["the service"], ["40:51"]], "Polar_expression":
[["ruined"], ["52:58"]], "Polarity": "Negative", "Intensity":
"Standard"}]},
```