NLU course projects lab 6

Davide De Martini (mat. 248445)

University of Trento

davide.demartini@studenti.unitn.it

1. Introduction

This report outlines the development of a model for opinion target extraction, that is a sub-task of Aspect Based Sentiment Analysis. The architecture is based on BERT, and it has to be fine-tuned for this specific task.

The dataset used is Laptop partition of SemEval2014 and the metrics used for the evaluation are precision, recall and f1 score.

2. Implementation details

In the paper by Hu et al. [1] the proposed method is a span based extract-then-classify framework. Instead of using a sequence tagging method, it extracts candidates by predicting the start and the end of a aspect in the sentence. However, in this implementation a sequence tagging method is used and this will affect the results as stated by Hu et al. [1].

The model is composed by a BERT (Devlin et al. [2]) pretrained layer and a linear layer for the fine-tuning, since BERT was not pretrained for sequence tagging tasks. A dropout layer before the linear one is added for regularizing the model.

The dataset is composed by the sentences and the tagging scheme. For the tagging scheme, the total number of classes is 3: the padding, the O (*Other*) class and the T (*Target*) class. Since there are three classes, we can transform the tagging scheme into a vector composed by 0 if there is padding, 1 if the tag is 'Other' and 2 if the tag is 'Target'. The dataset offers two versions of the sentences: raw and preprocessed. For better training, the preprocessed version is preferred. This preprocessing changes some words from the original to make them easier for the neural network to learn.

The sentences have to be preprocessed in order to be used by the BERT model. First tokenizing them, second expanding the tag where a token is splitted into a subtoken. If the token generates subtokens the first one will have associated the right tag id, the others will have associated the id corresponding to the padding tag. The latter thing is done in order to maintain consistency between the length of the sequence and the tagging scheme. Last, a padding at the start and at the end of the tagging scheme is added in order to match the [CLS] and the [SEP] tokens (see table 3 for a clarification).

The loss criterion used is the cross entropy loss since the nature of the task. A weight for the loss is also applied because the dataset is pretty unbalanced. The weights for the classes are calculated from the training set and then passed to the cross-entropy.

Since the training is pretty difficult, an L2 regularizer is added in order to better optimize the loss function.

For the evaluation part, the data is preprocessed in order to clear the ground truth of the extra pad tokens added in the preprocessing phase. In parallel, also the hypothesis are cleared, not by deleting the padding tokens, but deleting the tag with the index corresponding to the one in the ground truth having the padding token (see table 1 for a clarification). This is done for evaluation purposes, as the last step is to pass the hypothesis and the reference to the evaluation function.

reference	O	O	T-POS	~~[PAD]~~	O
hypothesis	O	T-POS	[PAD]	~~O~~	T-NEG

Table 1: *The strikeout text is the removed item in the hypothesis and reference*

The evaluation function was slightly modified to fit the data, but the implementation remains the same. The only two changed function are:

- the one for counting the number of hits (number of times that the extracted aspect match the ground truth).
- the one for counting the number of aspects in the hypothesis and in the reference.

3. Results

As stated from Hu et al. [1], using a sequence tagging technique for this task leads to a huge search. This affected the performance of the model and the slightly good results are the evidence of it.

The experiment was run for 40 epochs, the optimizer used is AdamW with $2 \cdot 10^{-5}$ of learning rate, the pretrained model is 'bert-base-uncased'. The batch size used are: 16 for the training set, 32 for the validation and test set. The results can be consulted at table 2.

Metric	Score
Precision	0.81
Recall	0.68
F1 score	0.74

Table 2: *Result for opinion aspect extraction task*

4. References

[1] M. Hu, Y. Peng, Z. Huang, D. Li, and Y. Lv, "Open-domain targeted sentiment analysis via span-based extraction and classification," 2019.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

sentence	[CLS]	con	##s	:	screen	resolution	[SEP]
tagging	[PAD]	O	[PAD]	O	T-NEG	T-NEG	[PAD]

Table 3: *Process of fitting the length of the sentence with the length of the tagging sequence*