HW3: Coreference Resolution

Bota Duisenbay
La Sapienza University of Rome
`duisenbay.1849680@studenti.uniroma1.it`

1 Introduction

Coreference Resolution is a NLP task aimed at disambiguating expressions (pronouns in our case) in a sentence, linking them to the corresponding entities. It can be divided into three steps as following: finding ambiguous expressions (pronoun identification), identification of all candidate entities (entity identification), and finally, determining to which entity the expression refers to (pronoun resolution). In this task, the last step was implemented as a classification problem as in ProBERT by (Attree, 2019).

2 Dataset

The dataset given for this task consists of the 2998 texts and 454 texts for training and validation. A text on average consists of 4.36 sentences and 77.7 words.

The data set is given as

- id
- text
- pron   entity_A   entity_B
- p_offset   offset_A   offset_B
- is_coref_A   is_coref_B

For pronoun resolution in the dataset, there are predefined two candidate entities. This allows us to formulate a task as a classification tasks with classes 'Candidate A', 'Candidate B' and 'Neither'. The true label distribution of the classes is summarized in the table 1 and the pronoun distribution in figures 1

	Cand. A	Cand. B	Neither	Total
Train	1330	1353	315	2998
Valid	187	205	62	454

Table 1: Class Distribution for Pronoun Resolution

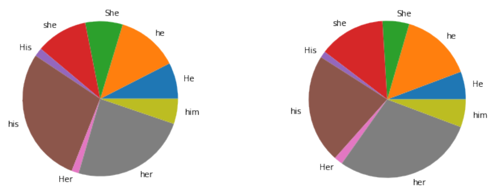

Figure 1: Pronoun distribution in training (left) and validation sets (right)

2.1 Pre-processing

As in ProBERT candidate entities and pronouns were enclosed by mention tags to implicitly indicate the position of labels in a text as in the example sentence below:
"...Jose de Venecia III, son of House Speaker <A>Jose de Venecia Jr<A>, alleged that <B>Abalos<B>offered <P>him<P>US $10 million to withdraw his proposal on the NBN project."

In the example sentence above, the pronoun "his" is enclosed with <P>, while the candidate entities "Jose de Venecia Jr" and "Abalos" with their corresponding tags <A>and <B>

The pronoun to be resolved and the candidate entities are given along with their (character) positions in a text. Therefore, pronoun and entity positions in a text must be represented in the final labeling as well to avoid confusion in case of the existence of the identical word in the same text. So, the pronoun and words of each entity must be aligned with their original positions after addition of mention tags and being tokenized.

The third class 'Neither' is assigned to a sample if both entities (is_coref_A and is_coref_B) are given as false.

3 Model: ProBERT

The model architecture implemented for this task is ProBERT by (Attree, 2019).

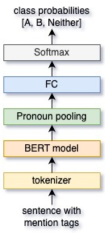

Figure 2: ProBERT model

learning rate	5e-6
weight decay	1e-4
betas	(0.9, 0.999)
epsilon	1e-06
batch size	8

Table 2: Training parameters

The sample text augumented with the mention tags is tokenized and feeds as input to the pre-trained BERT language model (Devlin et al., 2018). From the last layer of the model we do pronoun pooling, meaning, knowing the position of the pronoun, we extract token representation of the pronoun to be resolved. The pronoun token representation is passed to the linear layer, followed by softmax layers, to be classified into one of three classes. The architecture is summarized in the Fig. 2.

4 Experiments

The model was trained on colab gpus. Sequence classification models were used as language models. Language model and tokenizer used for training are 'albert-base-v1' [1] (Lan et al., 2019), bert-base-cased [2] (Devlin et al., 2018) and roberta-base [3] (Liu et al., 2019). AdamW optimizer (has better generalization performance than Adam) was used with parameters listed in the table 2. For regularization dropout of 0.1 is used at last layer of the transformer. Early stop with tolerance 2 was used to avoid overfitting.

5 Results

Three pre-trained language models used for experiments are albert base v1, bert base cased and roberta base. The results obtained from training different language models is summarised in the table 3. The best result obtained is overall 86.56% accuracy by roberta model received in 8 epochs. The confusion matrix reveals that accuracy of most present classes "A" and "B" are 86 and 91 % and they are most confused with each other, while the accuracy of less frequent class "Neither" is only 74%. This is a result of bias present in the dataset.

Model	Accuracy
albert base v1	78.63 %
bert base cased	72.03 %
roberta base	86.56 %

Table 3: ProBERT results for different language models

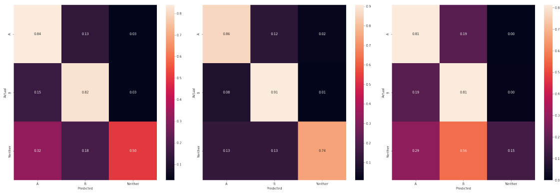

Figure 3: Confusion matrix for albert, roberta and bert model

6 Conclusion

Pronoun resolution in the given dataset was treated as a classification task and was solved using bert based model with ProBERT architecture. For the notion of model mention tags were introduced into the text to implicitly indicate each entity and pronoun. Different bert based language models were tested. The best result obtained is 86.56% by roberta model. It was suggested to reduce bias between classes in the dataset to improve the performance.

[1] https://huggingface.co/docs/transformers/model_doc/albert
[2] https://huggingface.co/docs/transformers/model_doc/bert
[3] https://huggingface.co/docs/transformers/model_doc/roberta

References

Sandeep Attree. 2019. Gendered ambiguous pronouns shared task: Boosting model confidence by evidence pooling. *arXiv preprint arXiv:1906.00839*.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut.

2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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