Homework 3: Conference Resolution

Mattia Pannone
pannone.1803328@studenti.uniroma1.it

Abstract

In natural language, one of the easiest things that can happen but difficult to identify in the absence of reasoning is the creation of misunderstandings due to ambiguity of the language itself if one is not able to attribute the right context. For this reason, Natural Language Processing involves developing algorithms and models that enable computers to understand, interpret, and generate human language in a meaningful way. There are several tasks and sub-tasks that focus on particular aspects of the language. In this homework, the relationship between pronouns and related entities in a text is explored, in particular by understanding, in the presence of an ambiguous pronoun, to which entity it may refer.

1 Introduction

Conference resolution is a task in Natural Language Processing (NLP) that involves resolving conferences in a text. In particular, given an ambiguous pronoun and two candidate entities, this task aims to determine which entity refers to the ambiguous pronoun. The one described above is actually the last step, called **entity resolution (3)**, of a series of sub-tasks to achieve the described goal. In this task two other sub-tasks are implemented: **ambiguous pronoun identification (1)** and **entity identification (2)**. In the first, the objective is to find which is the ambiguous pronoun (a text can contain multiple pronouns) that can refer to multiple entities (two in particular), while in the second it is precisely to find these entities to which it can refer, finally, as just said, the last step is to figure out which entity is right (if indeed it is one of the two, otherwise they are classified as false). The entire pipeline formed by these 3 steps is called **End-to-End (E2E) conferencing resolution**.

For example, in the sentence *"John went to Mike's store. He bought some milk."*, the pronoun "He" could refer to the entities "John" and "Mike". The conference resolution aims to identify such relationships within a given text, in order to make it semantically more meaningful and easier to understand, being able to understand that the pronoun refers to "Jhon".

2 Data Representation and Organization

The representation and organization of data, as well as the quantity and quality of data available, is essential for training a machine learning model and achieving satisfactory performance.

Having divided the homework into three subtasks, the data is organized in a similar but different way for each one. In particular, all the texts have been divided into tokens by two different tokenizers based on pre-trained models: the first based on the transformers useful for transforming the tokens into an input for the transformer itself, the other based on the python "spacy" library which allows you to keep track of the offset of each token in the text, along with its index. Furthermore, thanks to the latter which exploit POS-tagging, I identified all possible pronouns in each text, among which the model I trained identified the ambiguous pronoun. So for the first task related to the identification of the ambiguous pronoun, I concatenated to the representation of each token a flag set to one in case the aforesaid token was a pronoun, otherwise set to zero. In the second task for entity identification, I referred to the entity representation used in the recent Name Entity Recognition (NER) [paper] tasks called BIO encoding where each token is entity is identifiedidentified as Beginning (B) indicating the beginning of a named entity, Inside (I) indicating that the token is inside a named entity and Outdide (O) indicating that the token is outside of any named entities. In particular, in my representation to train the model, I associated the aforementioned classes with each token. In the representation instead I have associated a flag to indicate the position of the ambiguous pro-

noun. In the last task, having to perform a binary classification for each entity found, I deal with two representation approaches: in the first each input has been formatted as "[CLS] text [SEP] pronoun [SEP] entity [SEP]" where each text is been tokenized and used with all the tokens generated (even compound and/or unknown words that have been subdivided in turn into multiple tokens); in the second one I instead adopted the input "text [SEP] pronoun [SEP] entity" where for each text the first token of each tokenized word was taken.

3 Model Architecture

To carry out the 3 tasks I trained three different models on the related data with the related labels (the ambiguous pronouns for the ambguous pronoun identification task, the two entities involved and the true/false labels for each entity in the entity resolution task).

In all models I represented the data using the representation provided by a pre-trained model based on transformers, introduced in 2017 (Vaswani et al., 2017), and especially using BERT (Devlin et al., 2018). The transformer output was passed to a BiLSTM, a type of recurrent neural network capable of capturing the context of a text sequence by reading it in both directions (Hochreiter and Schmidhuber, 1997).

The output of the two layers just mentioned, in the tasks are passed to a linear classifier which will classify the tokens in the appropriate ways in the relative tasks, i.e. identification of a single token in task (1), classification in (B,I,O) of the token in the task (2) and binary classification of the output in task (3). However in the model of task (2) to improve the classification (as shown in tables [2]) a CRF layer (Condition Random Field) is added which is a standard model for predicting the most likely sequence of labels that correspond to a sequence of inputs (Lafferty et al., 2001) . Instead in task (3) the linear classifier does not receive the whole sequence, but only the sum of the features of the ambiguous pronoun and of the entity extracted from the text.

4 Training and Evaluation

The models were built and trained independently from each other, i.e. not in the form of a pipeline, but using the true labels provided in the dataset for each one.

After a process of hyperparameters finetuning, given the similarity of the models, some of those used are the same for each model such as the batch size of the data (32), the optimizer (Adam); the number of epochs (also using the early stopping technique) was instead 10 for tasks (1) and (3) and 50 for task (2), moreover the Mean Squared Error was used as loss function for identification (1) and binary classification (3) tasks while for task (2) it is used the loss provided by the CRF layer for training step and personalized loss that compare each elemnt predicted with the label for validation step. Moreover all the model used have been trained without performing the transformer fine-tuning, noticed that this not improved the performances.

All models were evaluated on the basis of accuracy (number of elements correctly identified / classified out of the total number), while in task (3) precision, recall and F1 were also added, being the third task the last of the pipeline, in which was important to see beyond the accuracy, also the ability of the model to avoid false positives and false negatives.

After executing the tasks ache in the pipeline, thus providing task (2) with the ambiguous pronouns identified by task (1) and task (3) with the pronouns and entities provided by the first two tasks, we can observe a drop in the presets of the entity resolution.

The image [1] show the loss curves for training and validation steps, the image [2a] show the confusion matrix for the entity resolution task. The tables [1,2,3] show some comparisons between some tests of the three different models of the three tasks.

5 More approaches on Entity Resolution

After training and evaluating the 3 models, I implemented two other approaches mainly aimed at performance comparison, for task (3).

In the first approach I used the same model using the transformers RoBERTa (Liu, Ott, Goyal, Du, Joshi, Chen, Stoyanov, 2019) and XML-RoBERTa (Conneau, Khandelwal, Goyal, Gu, Kirchhoff, Lukasik, Stoyanov, 2020) comparing the performances with BERT (table [4]) and then making an ensambling of the three models, however the performances did not improve.

In the second approach, based on a work combining BERT and SVM (Mohan Nair, 2022), I trained an autoencoder able to reduce the output (encoder) provided by the network made by BERT+BiLSTM and reconstruct it (decoder), subsequently I used

the encoder to reduce (to 300) the dimensionality provided by BERT(798) , to then provide its output with its labels to the binary classifier SVM, which is a widely used algorithm in machine learning and pattern recognition, using different parameters (the kernel in particular) for it. Performance hasn't improved as expected, however it's important to notice the difference when using different approaches. Figure [2b] show the confusion matrix for the performance with ensamblig the models.

6 Conclusions

The E2E conference resolution task, as seen, consists mainly of three identifiable subtasks such as ambiguous pronoun identification, entities identification and entity resolution. obviously improving the performance of each of the subtasks will also improve the performance of the final task. For example in task (1) one could try to take all the pronouns present and the context of the sentence, perhaps together with the position of the text in which they are found, classifying each one as ambiguous or not in task (2) one could integrate/improve a Name Entity Recognition model to identify the entities involved given a pronoun and the context, as well as the structure of the text; in task (3) the complexity of the model could be improved, as well as the organization of the data in order to better extract the context of the pronoun and of the entity to establish whether the latter is associated with the indicated pronoun. Last but not least there is also a diversified hyperparameter finetuning to improve the performance of each model.

References

- Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

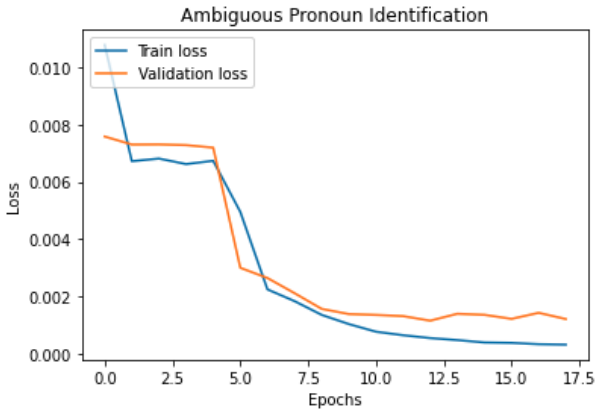
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30, 5998-6008.

- Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

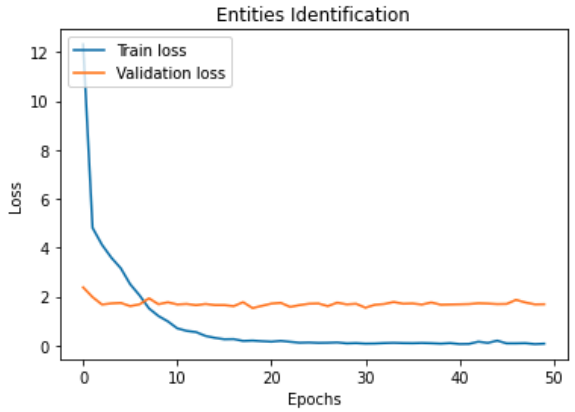
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

- Conneau, A., Khandelwal, G., Goyal, N., Gu, J., Kirchhoff, K., Lukasik, M., ... Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

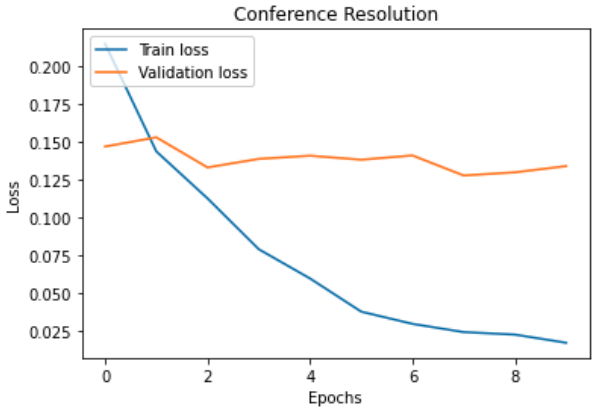
- Mohan, M., Nair, J. J. (2019). Coreference resolution in ambiguous pronouns using BERT and SVM. 2019 9th International Symposium on Embedded Computing and System Design (ISED), 10.1109/ISED48680.2019.9096245



(a)

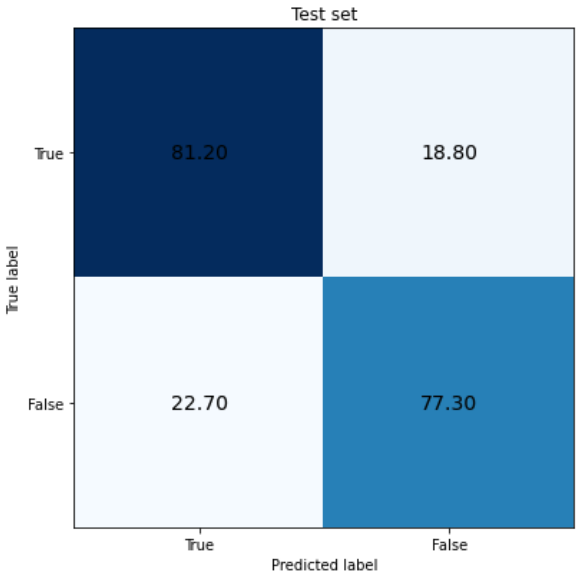


(b)

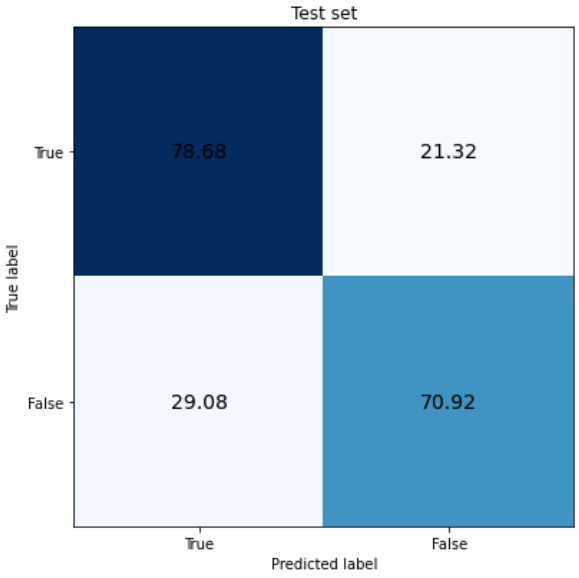


(c)

Figure 1: The above three images show the losses curves for training and validation steps for the three sub-tasks of Conference Resolution. **(a)** show the training of the ambiguous pronoun identification task, **(b)** show the training for the entities identification task, in this case the loss of train is given by the CRF layer, while the loss for validation is given by the cross entropy loss function. **(c)** show the training for the entity resolution task.



(a)



(b)

Figure 2: These images show the confusion matrix for the entity resolution task. **(a)** show the one for the model which reached best performance, **(b)** show the one for the ensambling of three different model. In both cases we can see, moreover the intensity of the blue color, the accuracy in predictions for true positives, true negatives, false positives and false negatives and the predictions reach a good score for the true labels with respect to false positives and negatives.

Model	Add features	Batch size	Bert fine-tune	epochs	Accuracy
Bert + 2 BiLSTM + 1 linear	Tag pronouns with '1'	32	no	50 (stopped at 18)	**85.46%**
Bert + 2 BiLSTM + 1 linear	Tag pronouns with '1'	32	yes	50 (stopped at 31)	**56.39%**
Bert + 2 BiLSTM + 1 linear	Tag pronouns with '1'	64/128	no	50 (stopped at 18)	**0%**

Table 1: Among those proposed, the model with the best performance is the one without fintuining of bert and with a batch size of 32. In the other two cases, especially by increasing the batch size, the model fails to learn.

Model	Add features	Bert fine-tune	epochs	Accuracy
Bert + 2 BiLSTM + 1 linear	-	no	50	**59.58%**
Bert + 2 BiLSTM + 1 linear + CRF	-	no	50	**64.21%**
Bert + 2 BiLSTM + 1 linear + CRF	-	yes	50	**0%**
Bert + 2 BiLSTM + 1 linear + CRF	Tag ambiguous pronoun with '1'	no	50	**64.10%**

Table 2: The accuracy for entities identification show the importance of a CRF layer in the classification of words sequence, it improves performance by almost +10%. Instead no improvements fintuning the transformer or tagging the ambiguos pronoun in the text.

Model	Data	Bert fine-tune	epochs	Accuracy	F1
Bert + 2 BiLSTM + 2 linear	mean of all word	no	10	**70.81%**	**69.64%**
Bert + 2 BiLSTM + 2 linear	pronoun+entity from last concatenated	no	10	**67.95%**	**63.40%**
Bert + 2 BiLSTM + 2 linear	pronoun+entity from middle in the text	no	10	**80.73%**	**78.84%**
Bert + 2 BiLSTM + 2 linear	pronoun+entity from middle in the text	yes	10	**77.64%**	**72.97%**

Table 3: For Entity resolution task it is computed Accuracy and F1 metrics. Both them improve if the classification is made on the sum of the pronoun and the entity taken by the representation in the middle of the text. For this task the finetuning of the transformer works, but it not improve the performances.

	BERT	RoBERTa	XML-RoBERTa	Ensambling
Accuracy	77.97%	73.79%	76.21%	75.33%
F1	75.43%	70.47%	72.80%	71.28%

Table 4: This table compare the Accuracy and F1 metrics between 3 different pre-trained transformers and their ensambling.