NLP Homework 3: Coreference Resolution

Marco Lo Pinto

1 Introduction

In the field of Natural Language Processing, Coreference Resolution is the task of automatically identify all the candidate entities in a text (mention proposal) and understand which pairs of mentions correspond to the same entity (mention linking). The task can be divided in three main parts: ambiguous pronoun identification (for simplicity, only one is ambiguous), entity identification and entity resolution (i.e. select the entity that corresponds to the identified pronoun).
The goal of this paper is to implement E2E Coreference Resolution, using the GAP Coreference Dataset (Webster et al., 2018).

2 Entity Resolution

2.1 Dataset preprocessing

In order to generalize the final model to evaluate multiple entities (and not just two, i.e. entity A and B), they are converted in a list of entities. The input phrase was processed differently for each tested model reported.

2.2 Simple trained sentence-level binary classification models

A sentence-level binary classification was tested: using a BERT-based model (fine-tuned in the process) and given to the model the ambiguous pronoun and a possible entity, it classifies if that entity corresponds to the pronoun (1) or not (0). Different configurations of the same model were tested: by focusing the output from the transformer to the entity using a mask, or concatenating it with the pronoun in order to improve results, or even exploiting the pooled output in order to add more context, as done in BertForSequenceClassification (huggingface), but the scores were pretty low (around 50%) and the training history was not encouraging (training and validation loss fluctuated greatly). A strategy with mention tags was tested, similar to the paper (Attree, 2019) but then discarded, due to the fact that the introduction of new tokens seem to confuse the model (probably if not trained with a great amount of data for a relatively long time).

2.3 Zero-shot binary sentence-level classification model

A zero-shot-classification pipeline was tried (Yin et al., 2019): in essence, given as input a sentence and some possible labels (custom made, that were never seen at training time), the model could construct a hypothesis and determine in which class the sentence should fall. The (pretrained) model used is BART (Lewis et al., 2019) after being trained on the MultiNLI (MNLI) dataset. For this particular task, the model received the sentence as premise and as hypotesis two dinstinct labels: "<pronoun> is <entity>" and "<pronoun> is not <entity>". The model performed well with simple phrases, but not with more complex ones, resulting in an accuracy of just 67%.

2.4 Question-Answering sentence-level binary classification model

Another test with a pretrained ready-to-use model was done using a fine-tuned RoBERTa-based model (Conneau et al., 2019) using the SQuAD2.0 dataset. The model received the sentence as context and the pronoun as question. Even though the model seemed to understand in some cases which was the referring entity, it didn't perform quite well with some input samples and it was discarded before evaluating it.

2.5 BERT-based entity resolution

Instead of outputting a single number for an input phrase, another way was tried: similarly to what done for the argument identification + argument classification model in my second homework (Lo

Pinto, 2022), one output for each (sub-)word was made. A BERT-based model is fine-tuned, and the input text is represented as [[CLS] sentence [SEP] pronoun [SEP]], in order to focus the attention mechanism on the target pronoun in the sentence. Then, a mask to isolate only the most probable entities is used (if the model receives it as input) and then passed onto a fully connected layer. This implementation generated better results, having almost 80% of accuracy, as we can see from table 1 and from the training history in figure 1. A clarification must be made: this model is general, meaning it can receive an arbitrary number of possible entities (or no entities at all, if trained properly).

3 Entity Identification

3.1 Dataset preprocessing

For this sub-task, a clever solution must be found: firstly, the dataset for this project is not useful, so another one was used, from the CoNLL03 NER dataset (Tjong Kim Sang and De Meulder, 2003).

3.2 BERT-based approach

The supposed model must take as input the sentence and output all possible entities that could refer to the pronoun. In order to do that, a similar approach to predicate identification from homework 2 was done: A BERT-based model is fine-tuned, and the input text is represented as [[CLS] sentence [SEP]]. Three possible labels are generated for each word: B = beginning of an entity, I = continuation of that entity and O = no entity. The last four layers are summed and passed onto a fully-connected layer. The model was trained on a small part of the CoNLL03 NER training dataset (around 40'000 sentences) and, as we can see from table 2, the evaluation with the dev dataset generated around 97% of f1-score using the seqeval library (Nakayama, 2018).

4 Entity Identification + Resolution model

Two main strategies were tested for the Entity Identification + Resolution model, using the evaluation system proposed for this homework: either use the proposed approach for Entity Resolution also in Entity Identification (i.e. without using the entities mask) or to use the Entity Identification model in combination with the Entity Resolution as a pipeline. As we can see from table 5, the best choice is the pipelined model.

5 Ambiguous Pronoun Identification

5.1 Dataset preprocessing

The dataset is preprocessed in a similar way as in the Entity Resolution part.

5.2 BERT-based approach

Similarly to what done in the Entity Identification part, a BERT-based model is fine-tuned, and the input text is represented as [[CLS] sentence [SEP]]. Then, a mask to isolate only the most probable pronouns is used and then passed onto a fully connected layer. The results are reported on table 3. The training history is reported on figure 3

6 Final model

The final model is composed of the Ambiguous Pronoun Identification on top of the Entity Identification + Resolution model. The score for this part is reported on table 4.

7 Results and Conclusions

The confusion matrix for the Entity Resolution model is in figure 4. The model seems to predict an entity as coreference even if that's not the case. This could be because of the unbalanced dataset (in 2684 samples there is the coreferenced entity, while only in 315 neither of them is the coreference).

References

Sandeep Attree. 2019. Gendered ambiguous pronouns shared task: Boosting model confidence by evidence pooling. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 134–146, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

huggingface. BERT huggingface. Accessed 2022-07-14.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Marco Lo Pinto. 2022. NLP homework 2: Semantic Role Labeling.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *CoRR*, abs/1909.00161.

Entity Resolution (model3)	**accuracy**
simple binary-output	0.5242
zero-shot based	0.6729
question-answer based	X
BERT-based + fc	0.7952

Table 1: Summary of architectures and their scores on the dev set for Entity Resolution. The "X" indicates that the model was discarded before evaluation.

Entity Identification (model2)	**f1-score**
BERT-based + fc	0.9718

Table 2: Summary of architectures and their scores on the dev set for Entity Identification.

Amb. Pron. Iden. (model1)	**accuracy**
BERT-based + fc	0.8898

Table 3: Summary of architectures and their scores on the dev set for Ambiguous Pronoun Identification.

E2E Coreference Resolution	**accuracy**
final model (BERT-based)	0.7137

Table 4: Summary of architectures and their scores on the dev set for E2E Coreference Resolution.

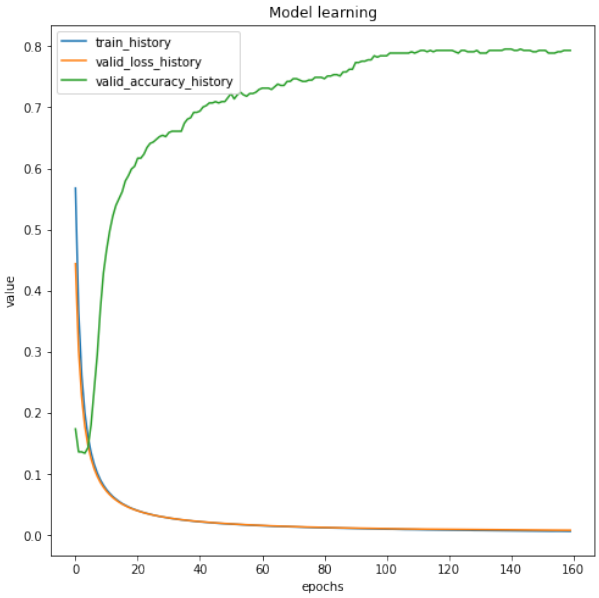

Figure 1: BERT-based for entity resolution training history.

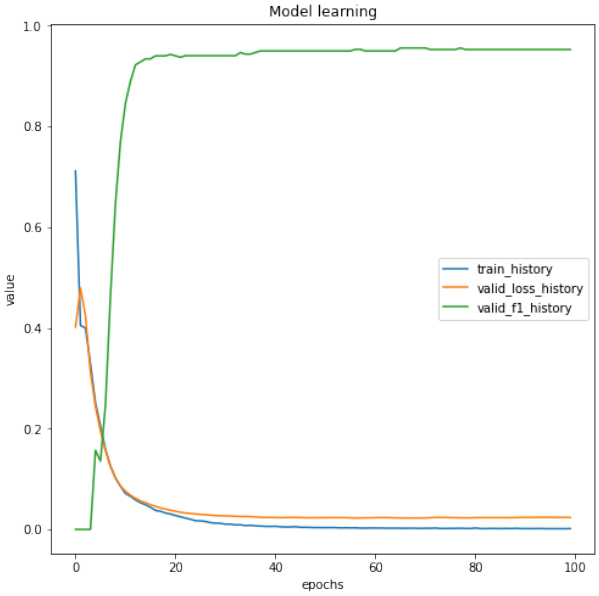

Figure 2: BERT-based for NER training history.

Entity Iden + Res (model23)	**accuracy**
NER-model + BERT-based	0.7950
only BERT-based	0.6541

Table 5: Summary of architectures and their scores on the dev set for Entity Identification + Resolution, evaluated via the script used for this homework.

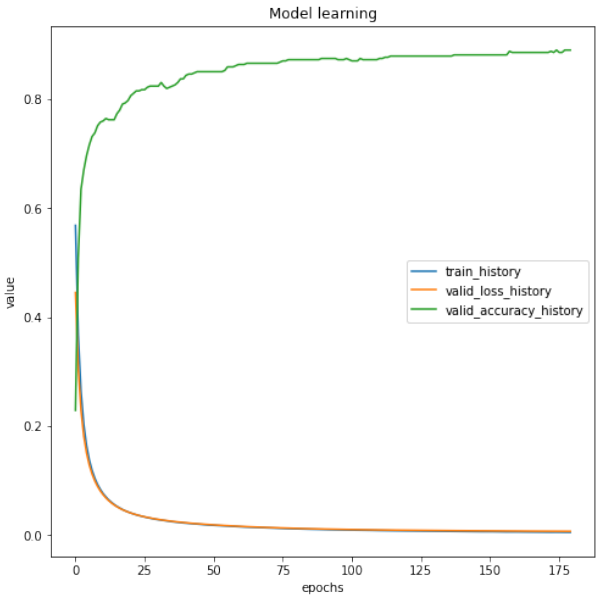

Figure 3: BERT-based for ambiguous pronoun identification training history.

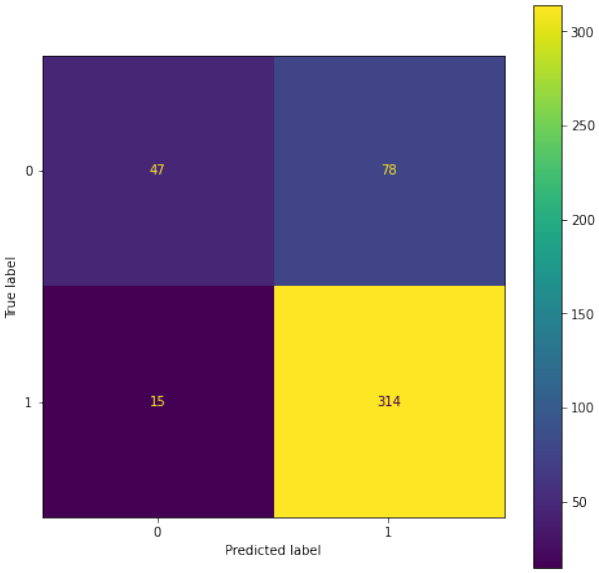

Figure 4: BERT-based for entity resolution confusion matrix.