SAPIENZA
UNIVERSITÀ DI ROMA

Homeworks presentation

“Natural Language Processing” course, a.y. 2021/22

Professor: Roberto Navigli

Candidate: Alessio Palma, palma.1837493@studenti.uniroma1.it

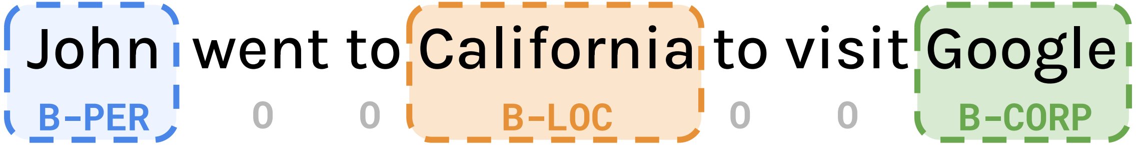
Named Entity Recognition - NER

Homework 1

The task and the dataset

NER is the task of automatically identifying named entities in a text and classifying them into some given categories, which depend on the chosen dataset. F1 score is the evaluation metric.

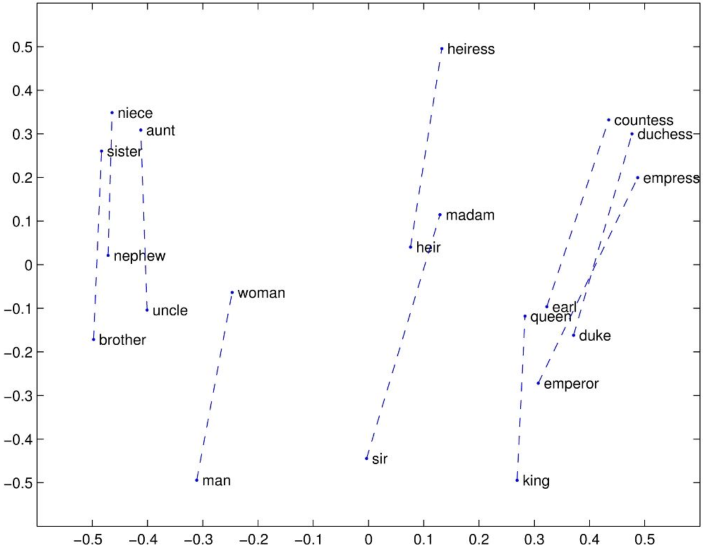
In our case we have 7 classes: **PER**son, **LOC**ation, **GR**ou**P**, **CORP**oration, **PROD**uct, **C**reative **W**ork and **O**ther. **BIO** tagging scheme is used to handle text spans that compose a named entity.

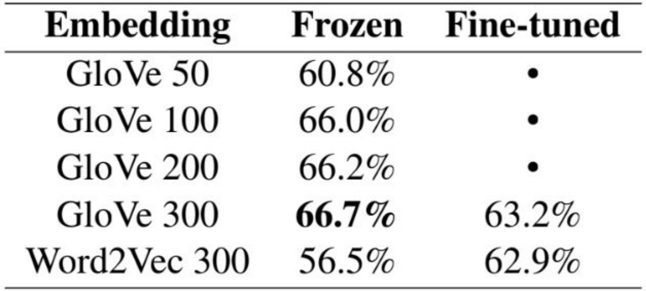


John	went	to	California	to	visit	Google
B-PER	O	O	B-LOC	O	O	B-CORP

The model - word embeddings

Each word needs to be transformed into a meaningful numerical representation, hence an embedding layer is used for this purpose. These word embeddings are **pretrained** on very large datasets, so they can better capture the semantic and syntactic relationships between words.





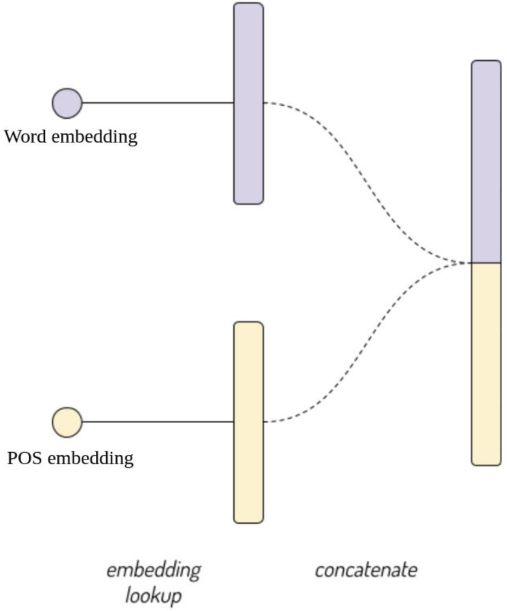
Embedding	Frozen	Fine-tuned
GloVe 50	60.8%	•
GloVe 100	66.0%	•
GloVe 200	66.2%	•
GloVe 300	**66.7%**	63.2%
Word2Vec 300	56.5%	62.9%

GloVe is the winner!

The model - POS embeddings

Part Of Speech (POS) Tagging is a main task in NLP that can often improve many downstream tasks. **Concatenating POS embeddings** to the word embeddings is useful because of the intuition that some words are most likely to be named entities with respect to other ones.

Stanza



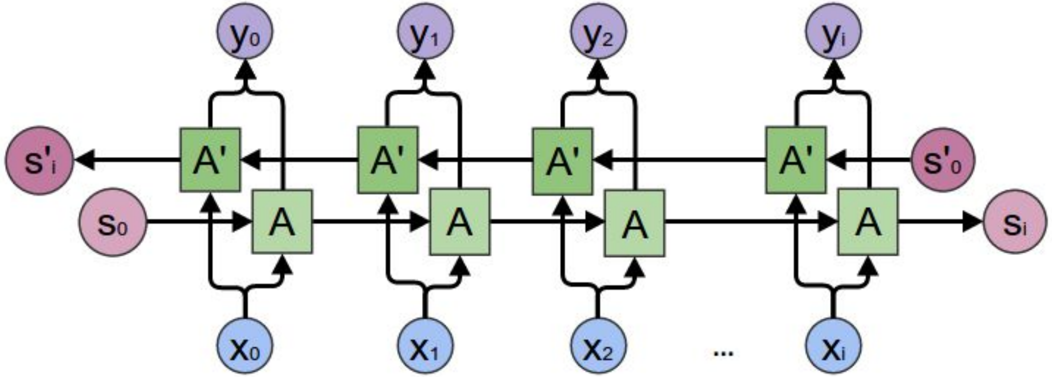
The model - BiLSTM

LSTM:

- + mitigates the vanishing gradient problem of RNNs;
- + allows for longer-term relationships;
- - only processes the input forward from left to right.

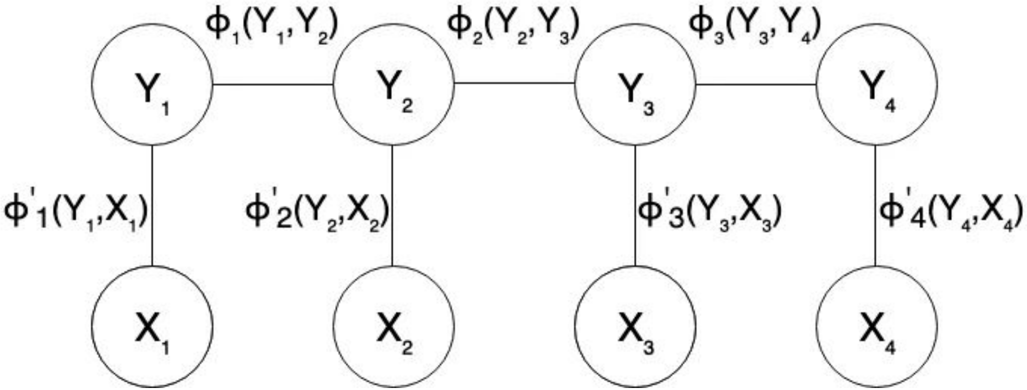
BiLSTM:

- + processes the input sequence both in forward and reverse order;
- - doubles the number of parameters.



After the BiLSTM there is a linear classification layer.

The model - CRF

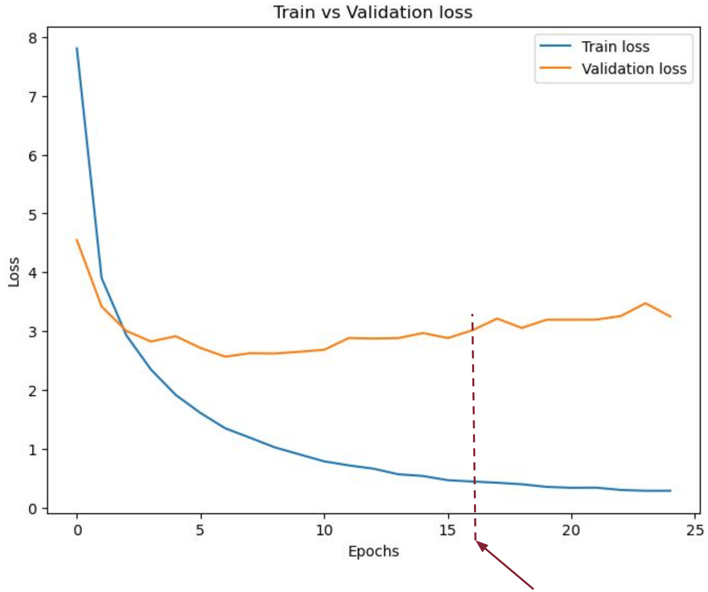


Conditional Random Fields (CRF) are a special case of Markov Random Fields that can learn constraints given the predictions at previous timesteps.

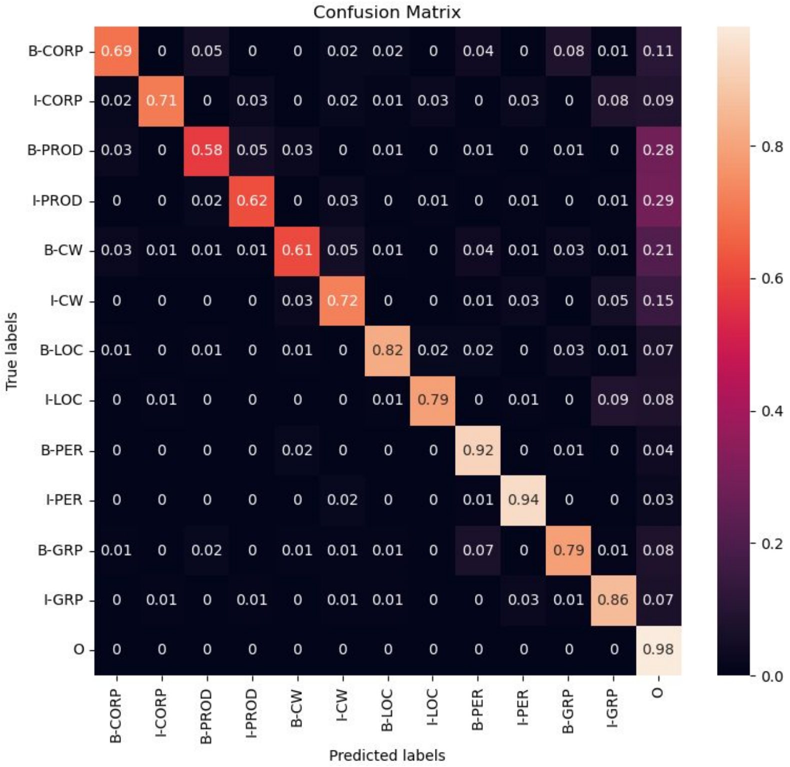
A CRF added on top of the linear layer forces the model to focus on producing **correct sequences of labels rather than correctly classifying individual tokens**.

SAPIENZA
UNIVERSITÀ DI ROMA

Results



Early stopping at epoch 24, best checkpoint saved at epoch 16. F1 validation score of **70.9%**.



	B-CORP	I-CORP	B-PROD	I-PROD	B-CW	I-CW	B-LOC	I-LOC	B-PER	I-PER	B-GRP	I-GRP	O
B-CORP	0.69	0	0.05	0	0	0.02	0.02	0	0.04	0	0.08	0.01	0.11
I-CORP	0.02	0.71	0	0.03	0	0.02	0.01	0.03	0	0.03	0	0.08	0.09
B-PROD	0.03	0	0.58	0.05	0.03	0	0.01	0	0.01	0	0.01	0	0.28
I-PROD	0	0	0.02	0.62	0	0.03	0	0.01	0	0.01	0	0.01	0.29
B-CW	0.03	0.01	0.01	0.01	0.61	0.05	0.01	0	0.04	0.01	0.03	0.01	0.21
I-CW	0	0	0	0	0.03	0.72	0	0	0.01	0.03	0	0.05	0.15
B-LOC	0.01	0	0.01	0	0.01	0	0.82	0.02	0.02	0	0.03	0.01	0.07
I-LOC	0	0.01	0	0	0	0	0.01	0.79	0	0.01	0	0.09	0.08
B-PER	0	0	0	0	0.02	0	0	0	0.92	0	0.01	0	0.04
I-PER	0	0	0	0	0	0.02	0	0	0.01	0.94	0	0	0.03
B-GRP	0.01	0	0.02	0	0.01	0.01	0.01	0	0.07	0	0.79	0.01	0.08
I-GRP	0	0.01	0	0.01	0	0.01	0.01	0	0	0.03	0.01	0.86	0.07
O	0	0	0	0	0	0	0	0	0	0	0	0	0.98

Many other details…

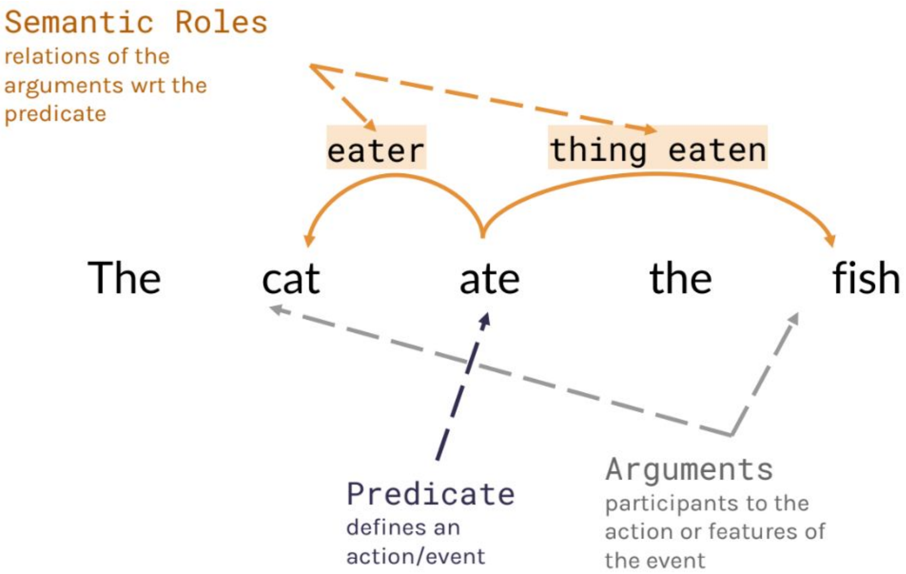
- Extensive random search over model's hyperparameters;
- Adam optimizer;
- Learning rate: 0.001;
- L2 regularization parameter: 1e-5;
- Loss function: CE when not using CRF, otherwise NLL;
- TorchCRF library used;
- Number of epochs: 60;
- Epochs of patience: 9;
- Batch size: 32;
- Saved the best checkpoint based on validation F1 score;
- Dropout of 35% after each layer.

Semantic Role Labeling - SRL

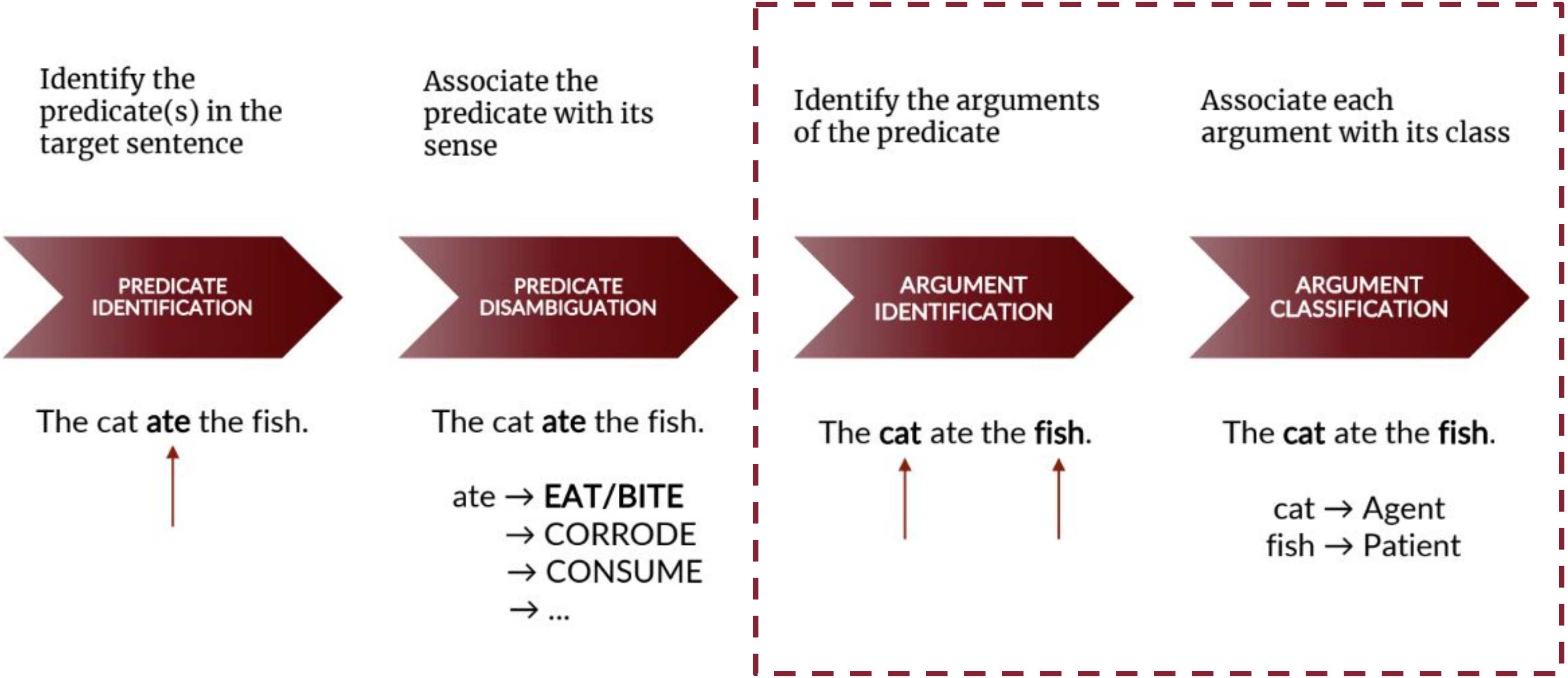
Homework 2

The task

SRL is the task of assigning labels to **arguments** in a sentence, indicating their **semantic role** with respect to a **predicate**. Informally, the task of addressing “Who did What to Whom, How, Where and When?”. F1 score is the evaluation metric.


Semantic Roles
relations of the arguments wrt the predicate
eater
thing eaten
The
cat
ate
the
fish
Predicate
defines an action/event
Arguments
participants to the action or features of the event

The pipeline



Identify the predicate(s) in the target sentence

PREDICATE IDENTIFICATION

The cat **ate** the fish.

Associate the predicate with its sense

PREDICATE DISAMBIGUATION

The cat **ate** the fish.

ate → **EAT/BITE**
→ CORRODE
→ CONSUME
→ ...

Identify the arguments of the predicate

ARGUMENT IDENTIFICATION

The **cat** ate the **fish**.

Associate each argument with its class

ARGUMENT CLASSIFICATION

The **cat** ate the **fish**.

cat → Agent
fish → Patient

The dataset

The dataset is **UniteD-SRL**[1], it contains 3 languages with parallel sentences and predicates are given. Predicate senses and semantic roles are defined according to **VerbAtlas**[2]. If there is more than one predicate in a sentence then I just replicate the sentence for each predicate.

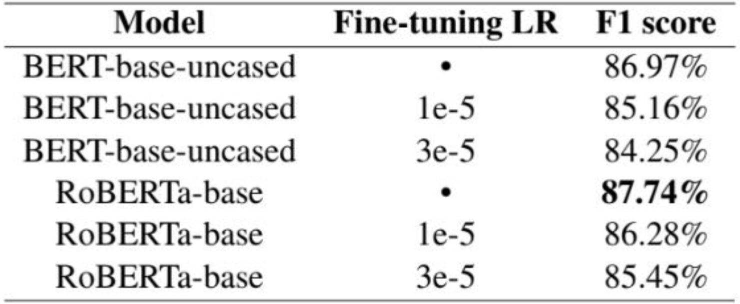
mandatory! →

	Train	Dev	Test
English	5501	1026	1027
Spanish	464	1026	1027
French	464	1026	1027

[1]: Rocco Tripodi, Simone Conia and Roberto Navigli: "UniteD-SRL: A unified dataset for span and dependency-based multilingual and cross-lingual Semantic Role Labeling." Proceedings of EMNLP 2021;
[2]: Andrea Di Fabio, Simone Conia, and Roberto Navigli. "VerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling." Proceedings of EMNLP-IJCNLP. 2019.

English model - contextualized word embeddings

The **contextualized embedding** for a word is important because it will be different based on the context of the sentence, as opposed to non-contextualized embeddings that will produce the same vector representation regardless of the context. In this way, different **senses of a word are not collapsed** into the main sense.



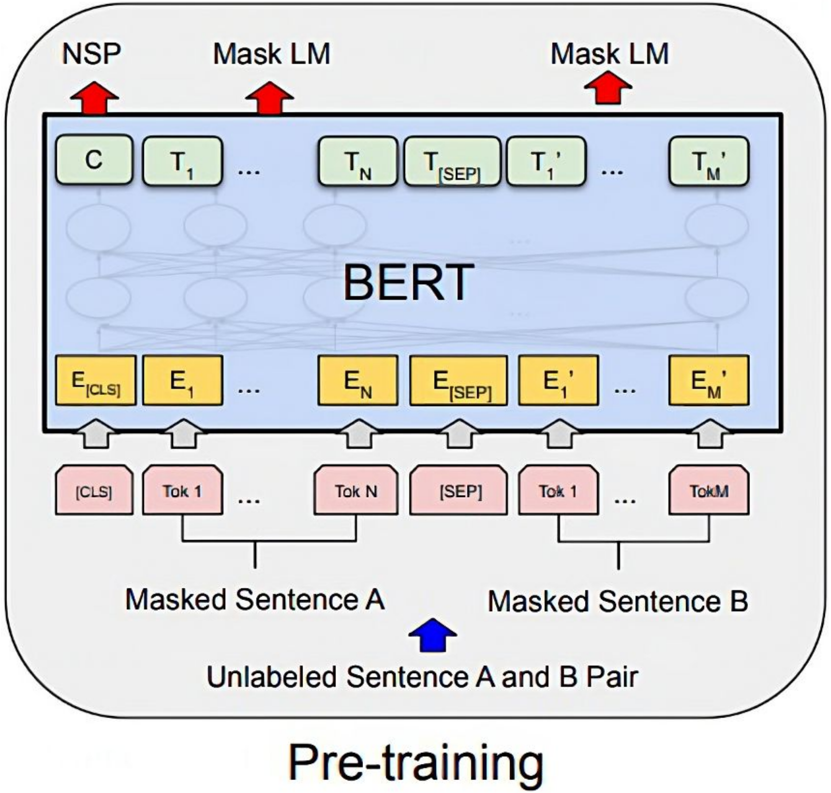
Model	Fine-tuning LR	F1 score
BERT-base-uncased	•	86.97%
BERT-base-uncased	1e-5	85.16%
BERT-base-uncased	3e-5	84.25%
RoBERTa-base	•	**87.74%**
RoBERTa-base	1e-5	86.28%
RoBERTa-base	3e-5	85.45%

RoBERTa wins!

English model - RoBERTa

RoBERTa is an **encoder-only Transformer** based on the same architecture as **BERT**, but trained on a lot more data (10x).

To obtain one embedding for each word, I averaged the last 4 hidden layers of RoBERTa and averaged the sub-tokens belonging to the same words.

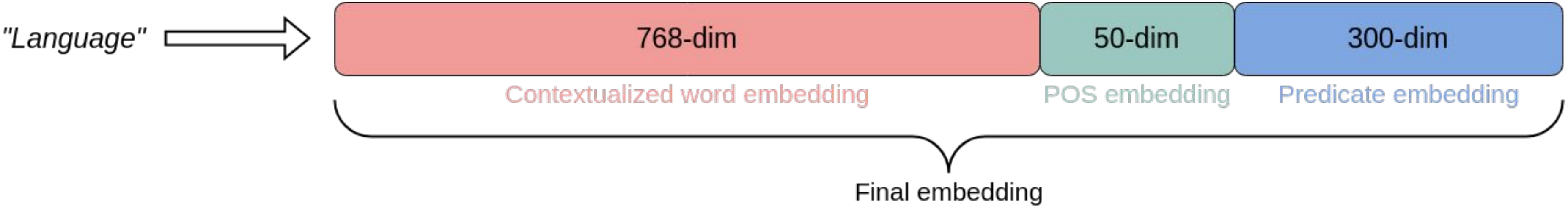


English model - POS and predicate embeddings

Concatenating POS and predicate embeddings to contextualized word embeddings is useful because of the intuitions that:

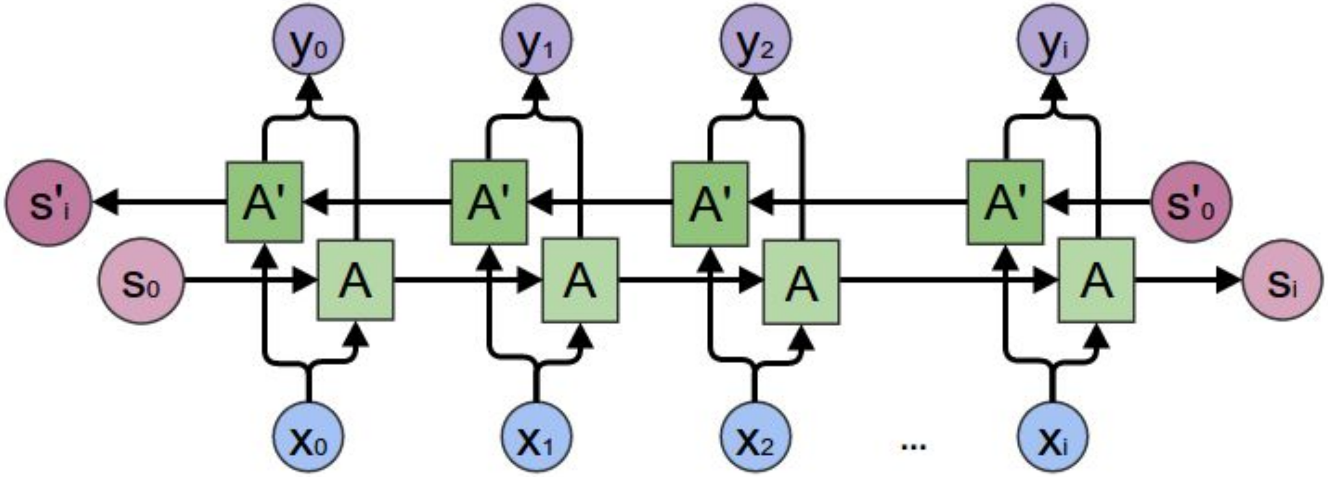
- some words are most likely to be semantic roles with respect to other ones;
- informing the system about where is the predicate in the sentence and which class it is, enables predicate-specific knowledge to be acquired.

Stanza



English model - BiLSTM

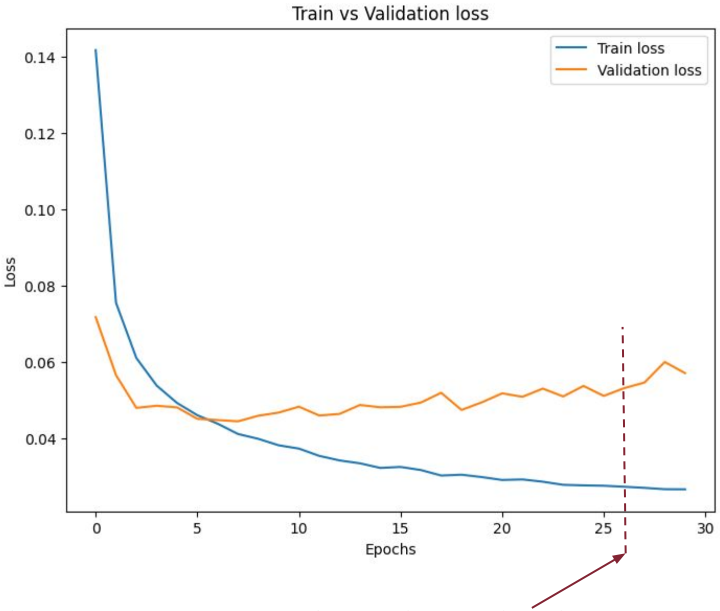
Final embeddings are fed to a BiLSTM.



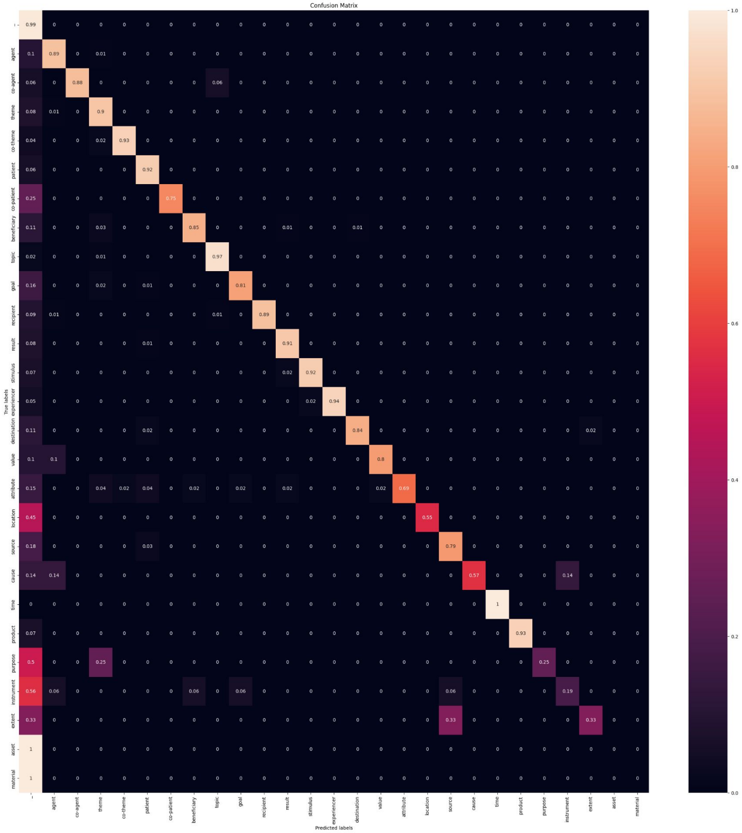
After the BiLSTM there is a linear classification layer.

SAPIENZA
UNIVERSITÀ DI ROMA

English model results



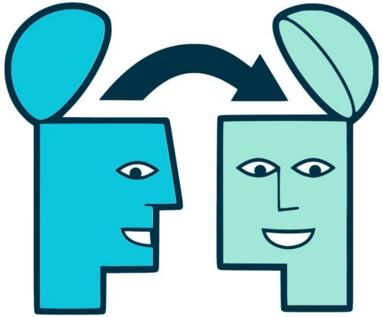
Early stopping at epoch 29, best checkpoint saved at epoch 26. Argument classification F1 validation score of **88.81%**.



Other languages

I compared three approaches based on the same architecture of the English model:

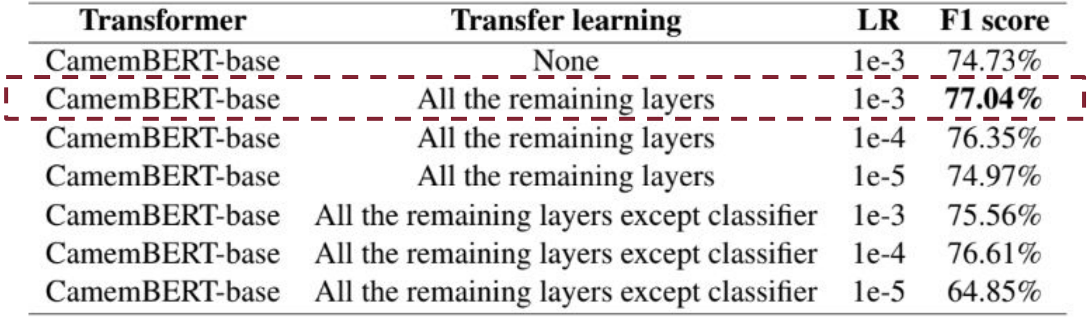
1) substitute the Transformer module with a language-specific one and retrain the whole model, which is randomly initialized;
2) substitute the Transformer module with a language-specific one, **transfer** weights from the English model for all the remaining layers and fine-tune;
3) substitute the Transformer module with a language specific one, transfer weights from the English model for all the remaining layers except the last linear layer and fine-tune.



2nd approach wins!

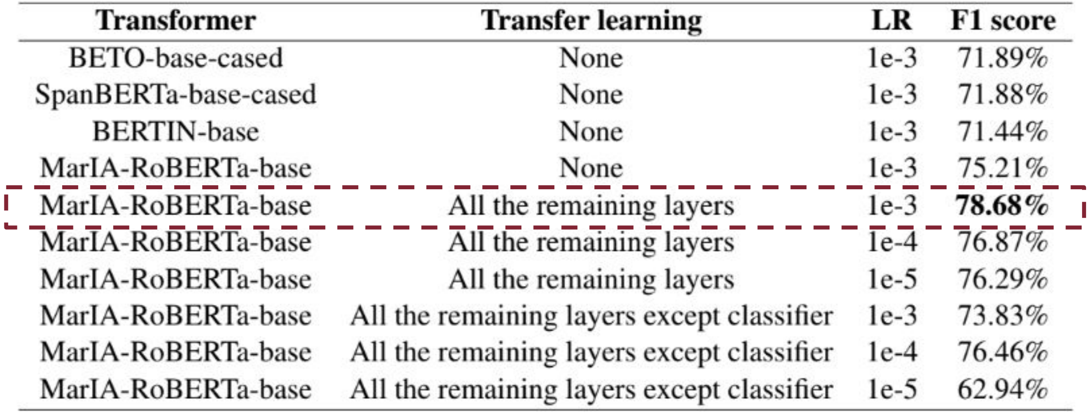
Other languages - results

French



Transformer	Transfer learning	LR	F1 score
CamemBERT-base	None	1e-3	74.73%
CamemBERT-base	All the remaining layers	1e-3	**77.04%**
CamemBERT-base	All the remaining layers	1e-4	76.35%
CamemBERT-base	All the remaining layers	1e-5	74.97%
CamemBERT-base	All the remaining layers except classifier	1e-3	75.56%
CamemBERT-base	All the remaining layers except classifier	1e-4	76.61%
CamemBERT-base	All the remaining layers except classifier	1e-5	64.85%

Spanish



Transformer	Transfer learning	LR	F1 score
BETO-base-cased	None	1e-3	71.89%
SpanBERTa-base-cased	None	1e-3	71.88%
BERTIN-base	None	1e-3	71.44%
MarIA-RoBERTa-base	None	1e-3	75.21%
MarIA-RoBERTa-base	All the remaining layers	1e-3	**78.68%**
MarIA-RoBERTa-base	All the remaining layers	1e-4	76.87%
MarIA-RoBERTa-base	All the remaining layers	1e-5	76.29%
MarIA-RoBERTa-base	All the remaining layers except classifier	1e-3	73.83%
MarIA-RoBERTa-base	All the remaining layers except classifier	1e-4	76.46%
MarIA-RoBERTa-base	All the remaining layers except classifier	1e-5	62.94%

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Many other details…

- Grid search over model's hyperparameters;
- Adam optimizer;
- Learning rate: 0.001;
- Loss function: Cross Entropy;
- Transformers Embedder library used;
- Number of epochs: 60;
- Epochs of patience: 9;
- Batch size: 32;
- Saved the best checkpoint based on validation F1 score;
- Dropout of 35% after each layer.

Coreference Resolution - CR

Homework 3

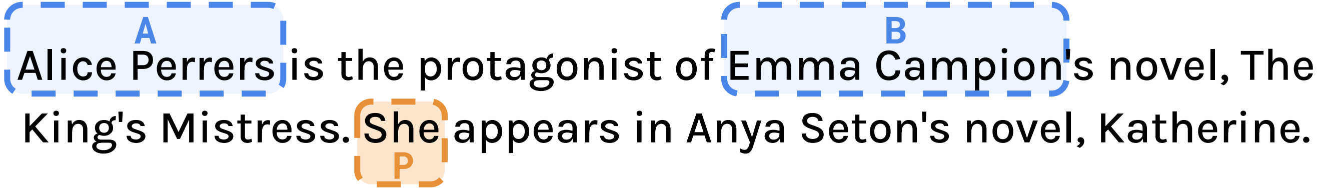
The task and the dataset

CR is the task of determining linguistic expressions that **refer to the same real-world entity** in a text. Accuracy is the evaluation metric.

It can be approached at three different complexity levels:

- End-to-End Coreference Resolution;
- Entity Identification and Resolution;
- Entity Resolution;

The **GAP dataset**[1] presents the Entity Resolution problem in a "gold-two-mention" format, phrasing it as a classification problem where the model must **resolve a given pronoun to either of the two given candidates or neither**. So the possible classes are: A, B, NEITHER.

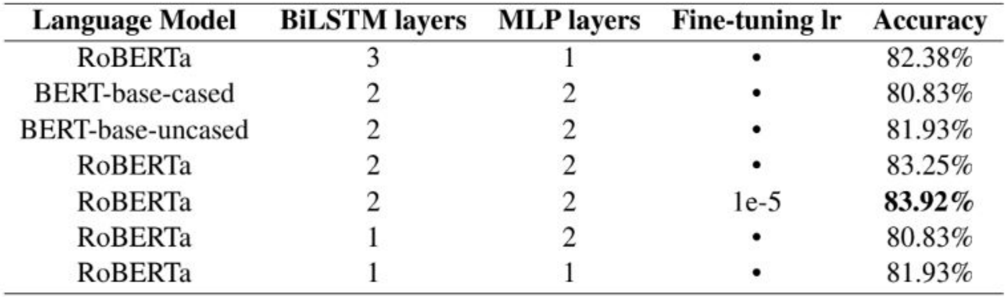


Alice Perrers is the protagonist of Emma Campion's novel, The King's Mistress. She appears in Anya Seton's novel, Katherine.

[1]: Webster, Kellie and Recasens, Marta and Axelrod, Vera and Baldridge, Jasoni: "Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns" Transactions of the ACL 2018.

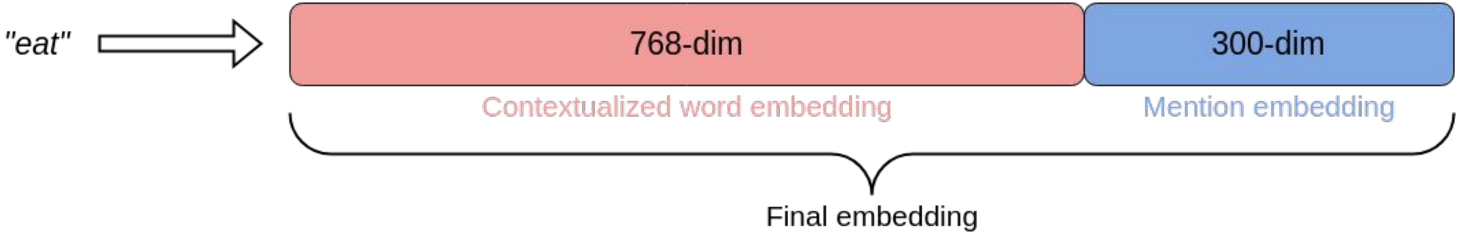
The baseline model - embeddings

(Fine-tuned) **RoBERTa** strikes again! To obtain a contextualized embedding for each word, I averaged the last 4 hidden layers of RoBERTa and averaged the sub-tokens belonging to the same words.



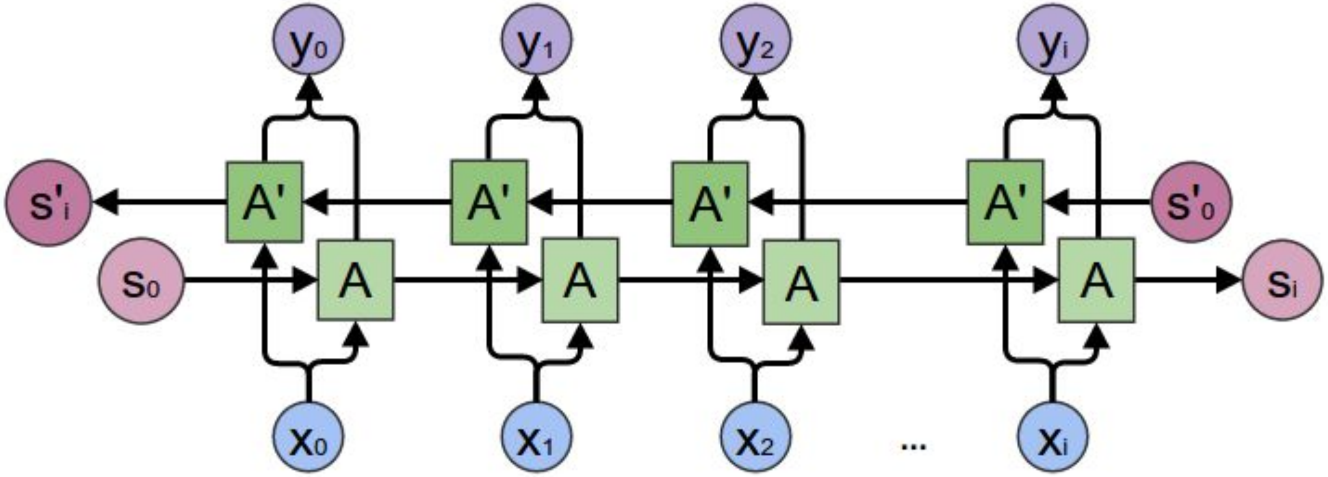
Language Model	BiLSTM layers	MLP layers	Fine-tuning lr	Accuracy
RoBERTa	3	1	•	82.38%
BERT-base-cased	2	2	•	80.83%
BERT-base-uncased	2	2	•	81.93%
RoBERTa	2	2	•	83.25%
RoBERTa	2	2	1e-5	**83.92%**
RoBERTa	1	2	•	80.83%
RoBERTa	1	1	•	81.93%

Concatenate to it a **“mention embedding”**, an embedding representing if the token is part of the pronoun, entity A, entity B or none of them. Useful to give the model a better understanding of which are the most important tokens in the sentence.



The baseline model - BiLSTM

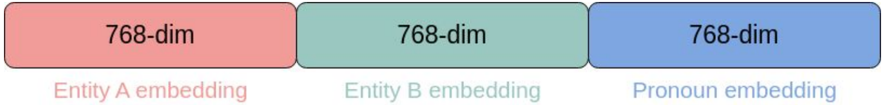
Final embeddings are fed to a BiLSTM.

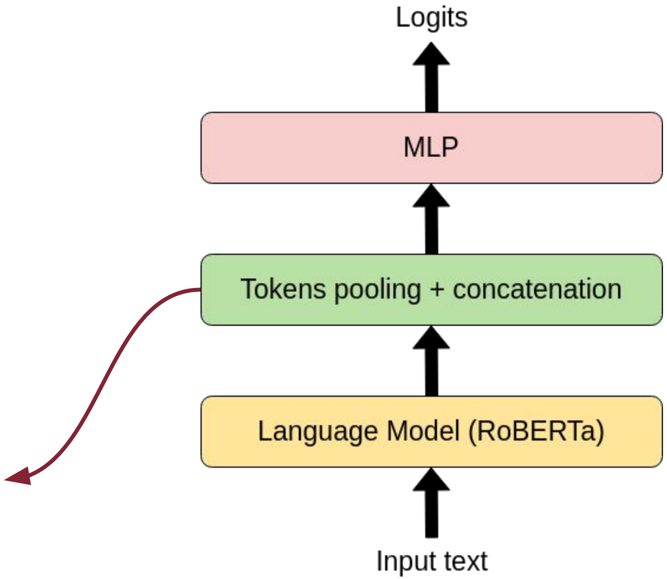


After the BiLSTM, I extract **only the features produced for the pronoun** token and fed it to a 2-layers MLP for classification.

CorefSeq[1]

(Fine-tuned) RoBERTa produces contextualized word embeddings, then I **concatenate the 3 separated embeddings** produced for entity A, B and the pronoun. This is fed to a 1-layer MLP for classification.

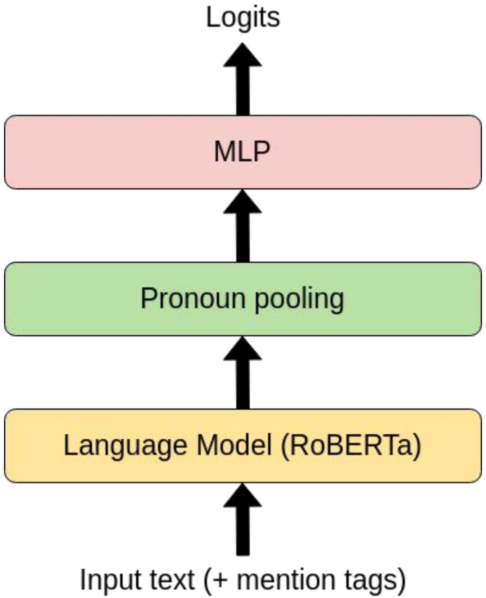




Simpler model, almost **+2.5%** improvement in accuracy!

[1]: Rakesh Chada. 2019. “Gendered Pronoun Resolution using BERT and an Extractive Question Answering Formulation”. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 126–133, Florence, Italy. Association for Computational Linguistics.

ProBERT[1]



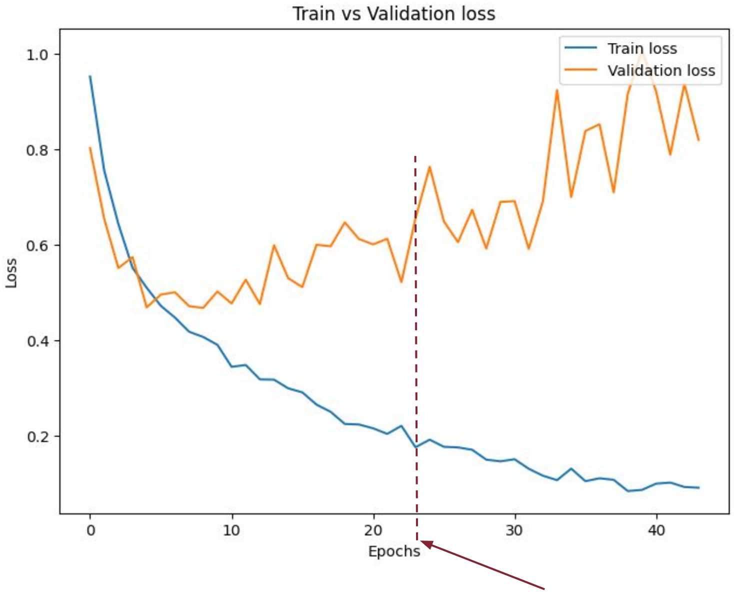
(Fine-tuned) RoBERTa produces contextualized word embeddings, then I extract the **embedding produced for the pronoun**. This is fed to a linear layer for classification.

Input text is augmented with **"mention level tags"** to capture positional informations implicitly:
*"**<A>** Bob Suter **<A>** is the uncle of **<B>** Dehner **<B>**.*
*   **<P>** His **<P>** cousin is Minnesota Wild's captain"*.

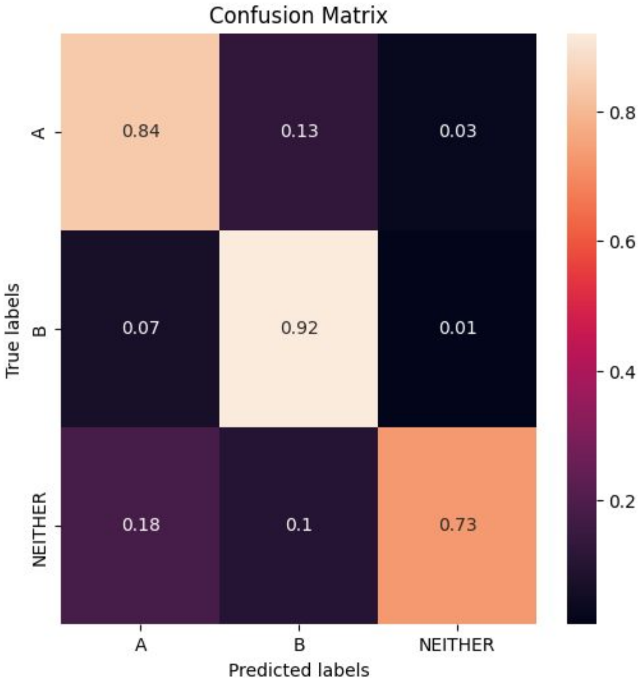
Even simpler model but the best, another **+1.5%** improvement in accuracy!

[1]: 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.

Results



Early stopping at epoch 44, best checkpoint saved at epoch 23. Validation accuracy of **89.86%**.



Many other details…

- Grid search over model's hyperparameters;
- Adam optimizer;
- MLP learning rate: 1e-4;
- MLP L2 regularization parameter: 1e-5;
- Transformer learning rate: 4e-6;
- Loss function: Cross Entropy;
- Transformers Embedder library used;
- Number of epochs: 70;
- Epochs of patience: 15;
- Batch size: 8;
- Saved the best checkpoint based on validation Accuracy score;
- Dropout of 35% after each layer.

Thank you for listening!