SRL Homework 2

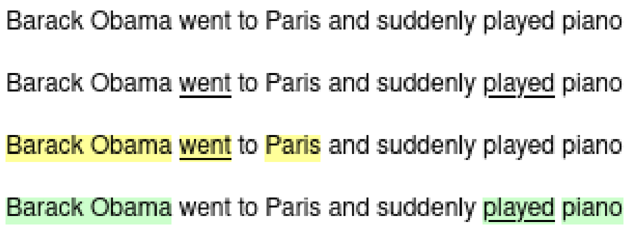
Francesco Peracchia
1906895
peracchia@studenti.uniroma1.it

Abstract

Semantic Role Labelling (SRL) is a fundamental task in natural language processing, using a BERT-based model for derive contextualized embedding, an architecture that combines multiple inputs information is proposed. Particularly this solution is inspired by (Di Fabio et al., 2019a), their approach is extended integrating additional embedding, and proposing a training procedure to increase Cross-lingual performance by transfer learning procedures between English,French and Spanish.

1 Introduction

SRL or Semantic Role Labelling is an fundamental task in NLP, the main goal of this task it to determine roles in a specific sentence, answering questions such as ”How,When,Where or Who did what to whom” determining the correspondent argument and label. A sentence always proposes entities and predicates, is then possible determine which elements of the sentence are entities and which are predicated, these processes are called argument and predicated identification. Afterwards since predicates can have different meanings we want to disambiguate those verbs attaching to them a proper contextualized meaning, then also arguments must be classified into different classes depending on their role into the sentence (Di Fabio et al., 2019b). Since many actors are involved in this process is more clear to divide the entire pipeline into smaller pieces and give them an order. In this work is proposed a SRL approach to identify and classify arguments give a certain predicated and the related meaning, then since this approach was originally applied to English, a transfer learning approach is proposed to increase results in French and Spanish dataset.



Barack Obama went to Paris and suddenly played piano

Barack Obama went to Paris and suddenly played piano

Barack Obama went to Paris and suddenly played piano

Barack Obama went to Paris and suddenly played piano

Figure 1: Predicates ”went” and ”played” are given, then we neet to identify arguments and classify them accordingly with the context.

2 Data processing

In this section is presented the data processing pipeline used to process English input, however it should be intended that the same procedure is also applied to French and Spanish. The following ordered list of actions is performed to generate our inputs information.

- Individual copies are created for each predicted.
- POS information is replicated for each copy.
- Predicated Flag indicator and predicated meaning embeddings for each copy with the related predicate.

Since we are only handling argument identification and classification is possible use the given identified predicate, hence counting the number of predicated present in the text we can create a number of text copy for each predicate, afterwards we can attach POS and predicate embeddings 1.

In order to exploit all the information given,are also used the disambiguate meaning of each predicated, and a positional flag are attached to to each copy 2.

Finally, Part of Speech (POS), predicate flag and predicate meaning flag are used to combine

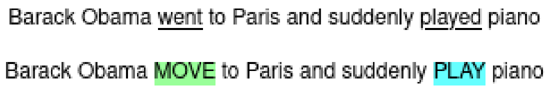

Barack Obama went to Paris and suddenly played piano
Barack Obama MOVE to Paris and suddenly PLAY piano

Figure 2: Identified predicates are successively classified into more then 300 labels, this information could be used in order to integrate more details to understand which is the role of each identified entity, exploiting the relation between predicate meaning and relative arguments.

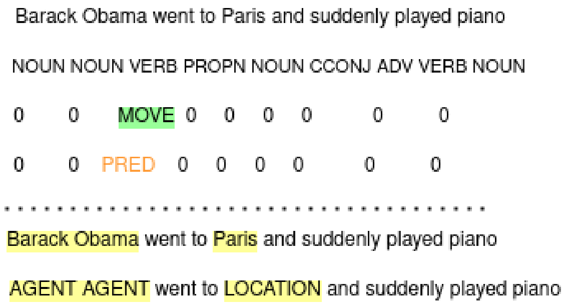

Barack Obama went to Paris and suddenly played piano
NOUN NOUN VERB PROPN NOUN CCONJ ADV VERB NOUN
0 0 MOVE 0 0 0 0 0 0
0 0 PRED 0 0 0 0 0 0
Barack Obama went to Paris and suddenly played piano
AGENT AGENT went to LOCATION and suddenly played piano

Figure 3: Searching for arguments related to predicate ”went”, requires that the desired predicate should be somehow indicated, this goal is reached by using 2 different flags, one flag is a simple ”predicate yes”, ”predicate no”, the other is integrating also the labels of the predicate, in this case ”MOVE”.

all the possible available information in order to create a pre-predicate contextual information that incorporate also these additional information. The final input to the model is the proposed at 3 4.

Since multiple predicated can be present into the text each predicated and its arguments are processed separately, 3 and 4 together propose the input for the original text input, different predicate have same input text, and POS, but different predicated flags.

In this pre processing phase we are assuming the all the possible arguments tag labels, and predicate meaning labels are known, however is possible that the explored training dataset do not contains all the possibilities. In this sense for all POS,predicate meaning and argument labels embedding, another additional tag , ”UNK” is used every time the presented words is classified with a not encountered tag during the training procedure.

3 Model

Since the main goal is to apply cross language learning, our model for English dataset is designed to improve transfer learning with French and Spanish languages rather than being optimized for English performances itself. The idea behind was to feed a MLP, with features extracted by 2 different branches, one branch it should be language specific, that means that the features extracted are completely different from language to language, while the other feature extractive branch should be generating features portable or almost portable between all the involved languages. Following this concept first ”BERT cased” was used to generate contextualized representation for L words, in parallel multiple embedding layers are used to generate different embeddings for POS, predicate flag and predicate disambiguate meaning. These information are feed into 2 different Bi-LSTM with the purpose of generation language depended and independent features since for French ad Spanish it will be used a different BERT (Devlin et al., 2018) and Tokenizer version. For all the languages BERT can generates outputs at multiple layers, rather than just collect the last we are summing the output of last 4 layers, and in case of sub words, only the first words is used and the other discarded.

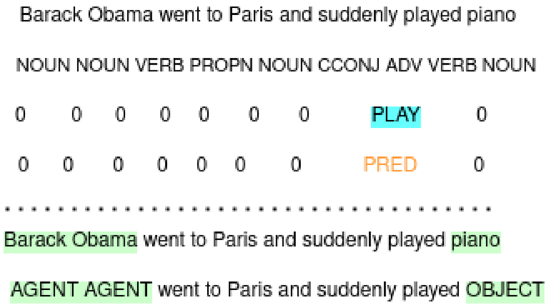

Barack Obama went to Paris and suddenly played piano
NOUN NOUN VERB PROPN NOUN CCONJ ADV VERB NOUN
0 0 0 0 0 0 0 PLAY 0
0 0 0 0 0 0 0 PRED 0
Barack Obama went to Paris and suddenly played piano
AGENT AGENT went to Paris and suddenly played OBJECT

Figure 4: Consequently, predicate ”played” generates other flag embeddings, for predicate position and predicate related label, founded after predicate disambiguation.

4 Training

Different embedding and hidden state size have been tried, however the final model use embeddings of size 768, 100, 50 and 32 respectively for BERT, POS, predicate disambiguate label, and predicate flag. The first branch receives in input a concatenated vector of size 950, while the portable branch an input of size 132 composed by POS and predicate flag embeddings. Then a 2 layers Bi-LSTM with internal dropout and a MLP that receives the output from the branches with 2 layers, Relu activation function and dropout. English model was trained at first with Adam optimizer and a scheduler that implements exponential decay of the learning rate, lr is initially set at 1e-3. Loss function choice rely on NLL that was anticipated by a logSotfMax layer. Then once the training was complete over the full English dataset, the model was saved and used for English test performances. In order to improve cross language exchange a successive fine tuning procedure was applied 6. An additional dropout layers was added into the language dependent branch, and for 2 epochs the model continue the training over the English dataset and forcing the model to rely more on the language independent branch. Then the model is saved, ”BERT cased” is exchanged with ”Bert multilingual-cased” at the embedding layer level, the second portable branch is frozen and goes accross all the languages, note that BERT was always use pre trained and with no gradient computation over its layers. In this way only the Bi-LSTM that involves new and previously unseen contextualized emdeddings from different languages is trained with the final MLP, these are the only parts the are leave to change and trained accordingly to the new dataset 7.

5 Results

In order to compare the effectiveness of this cross language approach the same model is evaluated across English French and Spanish dataset, metric reported are Precision, Recall and F1 score of the predictions over the entire test dataset. Since the label ”„ is the most common and also the only labels that does not find a correspondence into the task, ground truth labels are used to avoid those samples and their correspondent predictions into the metrics computation. Finally are reported both the results over French and Spanish languages with or whit out English transfer learning 2, for this experiments is used the configuration setup explained into the previous chapter and the strategy presented in these successive graphical representation 5 6 6. In order to contrast overfitting and keep performances over training and testing dataset close, different generalization techniques are used, especially different dropout layers between each Bi-LSTM, also before and inside the MLP are used for the final token classification. F1 score over the English data achieve a score of 0.771, early stopping with patience stops the training when the f1 score over the testing dataset starts decreasing. The fine-tuned model 6 trained over the previous one is used for transfer learning from English to French and English to Spanish. This approach results efficacy pushing up the f1 score obtained with the evaluation over the test dataset, for the model retrained after transfer learning then we achieve better results than results obtained without, again as before when the f1 score starts decreasing the training is interrupted with an early stopping procedure tested with different patience values. The same learning rate values is applied to both the approaches, and lead the model trained without transfer learning to overfit the training dataset only after 4 epochs for both Spanish and French, while for the alternative cross language approach, under same conditions except for the frozen layers encouter the same problem only after 14 epochs, and after achieving and higher score as reported in 2.

LANGUAGE	PRECISION	RECALL	F1
English	0.818	0.736	0.771
French	0.633	0.487	0.550
French*	0.567	0.316	0.406
Spanish	0.694	0.484	0.570
Spanish*	0.577	0.400	0.473

Table 1: Argument Classification, French*, and Spanish* results are obtained by the correspondent models with the same above proposed architecture and settings, just without transfer learning and frozen parts.

LANGUAGE	PRECISION	RECALL	F1
English	0.894	0.826	0.858
French	0.814	0.626	0.708
French*	0.923	0.515	0.661
Spanish	0.856	0.598	0.704
Spanish*	0.894	0.620	0.732

Table 2: Argument Identification

References

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

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019a. Simple bert models for relation extraction and semantic role labeling.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019b. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China. Association for Computational Linguistics.

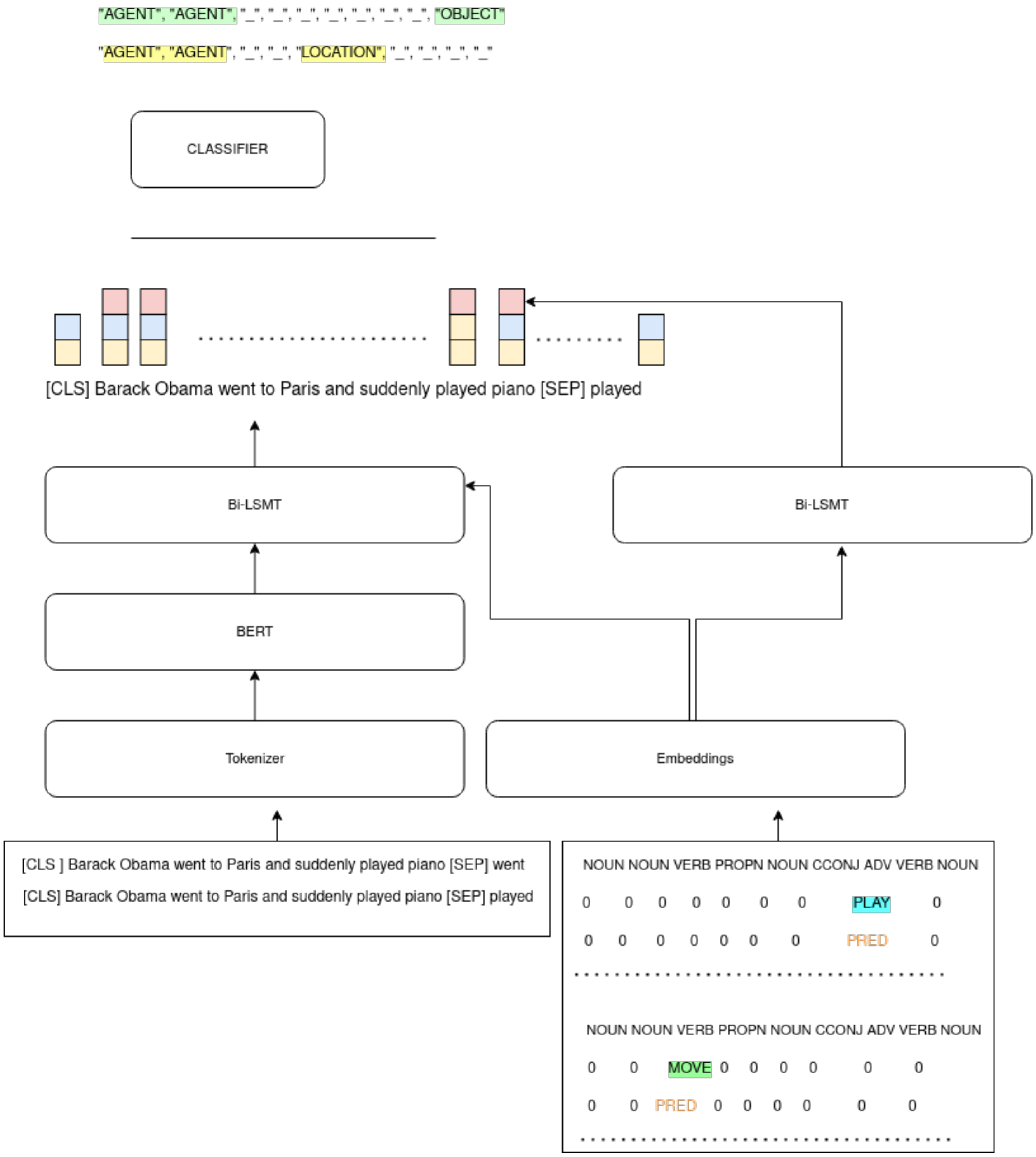

"AGENT", "AGENT", "_", "_", "_", "_", "_", "_", "_", "OBJECT"
"AGENT", "AGENT", "_", "_", "LOCATION", "_", "_", "_", "_"
CLASSIFIER
[CLS] Barack Obama went to Paris and suddenly played piano [SEP] played
Bi-LSMT
Bi-LSMT
BERT
Tokenizer
Embeddings
[CLS] Barack Obama went to Paris and suddenly played piano [SEP] went
[CLS] Barack Obama went to Paris and suddenly played piano [SEP] played
NOUN NOUN VERB PROPN NOUN CCONJ ADV VERB NOUN
0 0 0 0 0 0 0 PLAY 0
0 0 0 0 0 0 0 PRED 0
NOUN NOUN VERB PROPN NOUN CCONJ ADV VERB NOUN
0 0 MOVE 0 0 0 0 0 0
0 0 PRED 0 0 0 0 0 0

Figure 5: Argument identification and classification model for English, the related predicated is attached after the [SEP] token, POS, and predicated embeddings are then used to generated a complementary structural information, then finally BERT and the first Bi-LSTM complete the first branch while a second Bi-LSTM receives in input only the portable cross language information. Finally their output is concatenated, hence the relative predicate hidden state, the related token hidden state and the hidden state from the second branch are combined and feed to a MLP that works as a NER classifier.

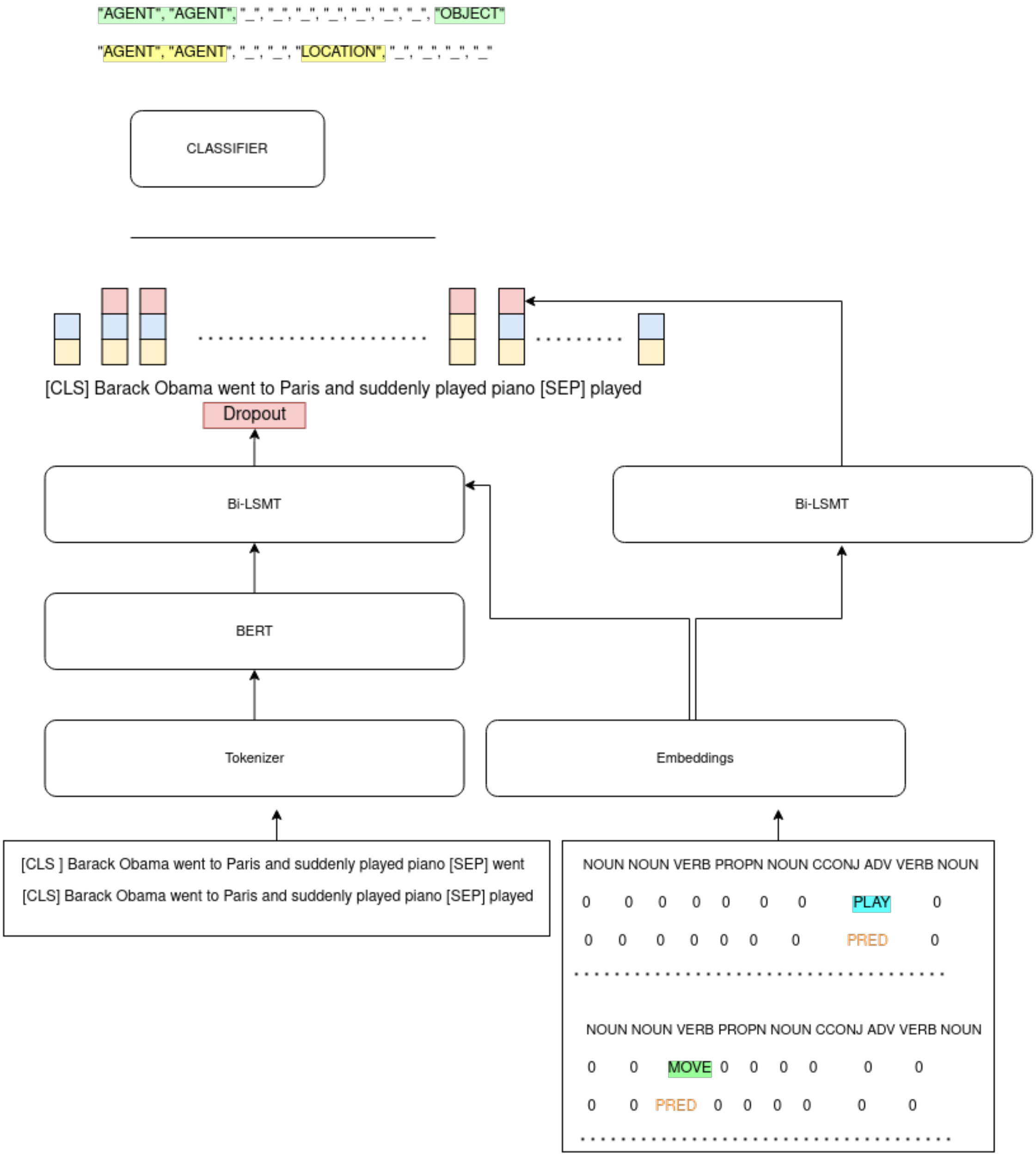

"AGENT", "AGENT", "_", "_", "_", "_", "_", "_", "_", "OBJECT"
"AGENT", "AGENT", "_", "_", "LOCATION", "_", "_", "_", "_"
CLASSIFIER
[CLS] Barack Obama went to Paris and suddenly played piano [SEP] played
Dropout
Bi-LSMT
Bi-LSMT
BERT
Tokenizer
Embeddings
[CLS] Barack Obama went to Paris and suddenly played piano [SEP] went
[CLS] Barack Obama went to Paris and suddenly played piano [SEP] played
NOUN NOUN VERB PROPN NOUN CCONJ ADV VERB NOUN
0 0 0 0 0 0 0 PLAY 0
0 0 0 0 0 0 0 PRED 0
NOUN NOUN VERB PROPN NOUN CCONJ ADV VERB NOUN
0 0 MOVE 0 0 0 0 0 0
0 0 PRED 0 0 0 0 0 0

Figure 6: In order prepare the model to a cross language training, must be forced to give more importance to the second branch, then another dropout layer is activated, in this case with probably $p = 0.6$ hidden states outputs from the first branch are set to 0, in this way the model is forced to extremize its predictions giving more importance to structural information, this is especially useful for argument identification, the obtained model is used as pre-trained base French and Spanish languages.

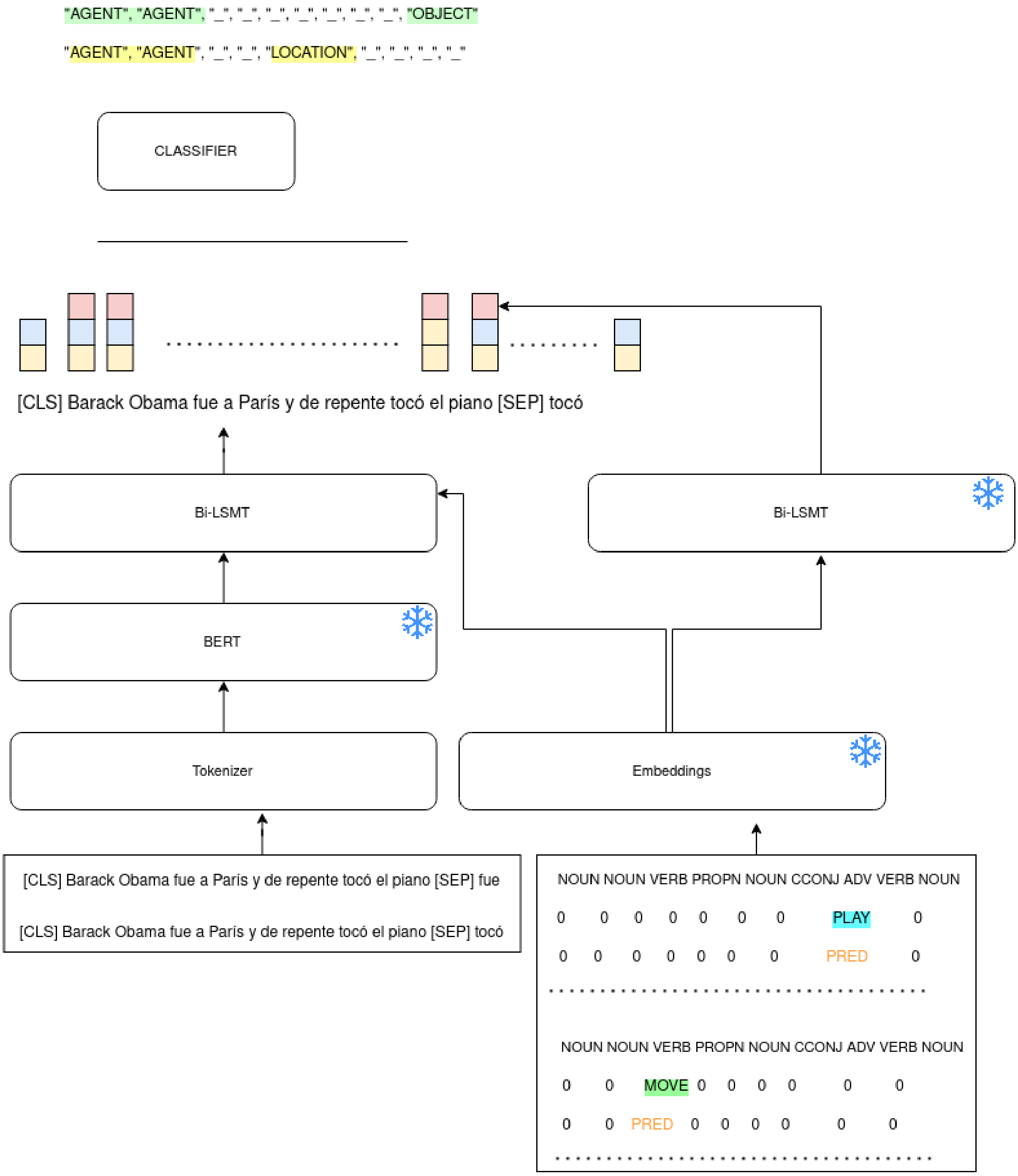

"AGENT", "AGENT", "_", "_", "_", "_", "_", "_", "_", "OBJECT"
"AGENT", "AGENT", "_", "_", "LOCATION", "_", "_", "_", "_"
CLASSIFIER
[CLS] Barack Obama fue a París y de repente tocó el piano [SEP] tocó
Bi-LSMT
Bi-LSMT
BERT
Tokenizer
Embeddings
[CLS] Barack Obama fue a París y de repente tocó el piano [SEP] fue
[CLS] Barack Obama fue a París y de repente tocó el piano [SEP] tocó
NOUN NOUN VERB PROPN NOUN CCONJ ADV VERB NOUN
0 0 0 0 0 0 0 PLAY 0
0 0 0 0 0 0 0 PRED 0
NOUN NOUN VERB PROPN NOUN CCONJ ADV VERB NOUN
0 0 MOVE 0 0 0 0 0 0
0 0 PRED 0 0 0 0 0 0

Figure 7: Cross language between English-French and English-Spanish is ensure by transfer learning, the previously fine tuned model is loaded for these new languages, however since the BERT model is now different the first branch must be re-trained, while the second that only depends on POS and predicate embeddings does not change across different languages. Then all the model except the Bi-LSTM is frozen.