**REPORT EX 4 NLP**

For this assignment, we need to implement a relation extraction system.

To do so, we received several sentences, with their tree parse, and the relation extracted from them.

We need, first of all, to read the files and to extract features that are important to us.

We chose to treat the relation : *Work\_For*.

**Use of Spacy**

To do so, we used the Spacy package. We used the Spacy Model « *en\_core\_web\_lg* ».

**Analysis and ameliorationאנחנו כתבנו פונקציה להשגת אנטיטיס וכאשר הרצנו ובדקנו את עצמנו זהינו שהצלחנו להוציא רק 60**

We create a function to get entities that we want to work with them for « Word\_For » and when we ran and checked it, we found out that we succeeded to get only 60% of them.

We tried different spacy models to get the maximum entities of the data.

We tried to ameliorate the parser, adding a constraint.

-With decomposition ‘-‘. We want to take into consideration the whole entity and not like two different entities.

It gave us a good results of identifications but not enough.

We tried to ameliorate even more :

After analysing the entities we saw that things like dot : ‘ . ’, ‘ ’s ’ at the end of entity ,‘the’ at the beggining of the entity or ‘the the’ in the beggining also reduce the results.

After several manipulations wih entitites we got to 96/109 entities in dev and train.

**Building the model**

When we built the model, we based on the form of the features and into sparse vector. For our first parameters, we tried to combine different features.

We played with different features like distance between the sentences, the words around and their parameters. It worked great on the training, but in the DEV evaluation and train, we got some failures. It didn’t identify the entities.

After figuring out that it didn’t work well, we add a Key Word vector. Indeed, the « Kill » relations are a lot similar with the « Work\_for » relation. Therefore, in the Key Word vector, we put words that belongs to the « kill » relation like « shoot », « assassinate », « death », « murder », « kill » etc… The model didn’t succeed to indentify them. And doing that, helped a lot !

We also add in the key words vector, words that can be useful for the « Work\_For » identification : « work », « head », « serve », « star », « perform » etc…

We played with different models : Logistic regression and SVM.

We tried to fit with one of our model to get the best one. So that the model will deal with the non balanced data, we added : Class\_weight ‘balanced’ that balance the data by changing the weights, thanks to the SKLearn Package.

The best model in Logistic regression is wih solver = ‘liblinear’, with penalty = ‘l1’ (regularization). But, it gave us a not great identification. Indeed, we got a lot of false positive examples.

The best model in SVM is with Kernel Linear, dregree 3 and c= 0.3. It gave us the optimal resulst.

We also tried to give different Kernels but it didn’t give us better results.

We have « Work\_For » sequences with entities that are not relevent. Ex : Source = Org, Label = Org, or Source = loc, Label = loc. It’s not helping to train our model. It turns out that it unbalance even more the data.

As a solution, we made a script with a description of what really should be sequence of « Work\_For » : the possible sequences for every label of work for.

Why ?

We have a non balanced data, and also sequences that don’t exist so we want them to be ‘False’ because we don’t want to take them into consideration. We want the model to learn only on the true ones.

The final model countains prediction of parse sentences, get every sequences of entities. For every entity, it checks : does this sequence exist in the possible sequences ?

**Hard and Soft evaluations**

We did two evaluations : hard and soft evaluations.

The hard one consist of finding label that fits the prediction.

The soft one :

To ameliorate our model, we softened the source and the target. In other words, if the source/target our model found is in annotation the source/target is True. And the opposite, if the annotation source/target is in the source/target our model found : True.

Example : target = Home Loan Bank of San Fransisco

Our model : Home Loan Bank of San Fransisco

Annotation : Home Loan Bank

In both case, it is suppose to identify.