Introductions to NLP – Relation Extraction 89680

Itai Mondshine 207814724

Itay Lotan 308453935

Explanation about our model

As we have learned in class the problem of relation extraction can be solved with different methods. Due to the amount we used have a machine learning approach with a classification model.

We chose the relation "Live_In" (the second one) beacuse we thought it will be easier to annotate place of living than place of working.

Also, in order to split the sentence to chunks and for NER we used the package Spacy.

<u>In order to improve the performance of our system, here are some interesting things that we did:</u>

- **Boolean vs multicalls approach -** at first we trained the model as a simple classification model, with two possible labels a output - is relation of type Live_In or not. Then, we tried to train the model as a multiclass calssifacation model, where all the possible relation types ('Kill', 'Located_In', 'Live_In', 'Work_For' and 'OrgBased_In') can be predicated by it (along with a special 'NO_RELATION' label for the negative cases). We learned that the multi-label approach yields better results, so we chose using it for our final model. Notice that in inference time, we only consider the 'Live_In' predictions as positive.

- **Classifier type -** for our classifier, we tried two options:
  - skealrn's SVM LinearSVC - similar to SVC with parameter kernel = 'linear'
  - Catboost classifier - a populaer tree based model

    Catboost achieved lower overfitting, together with better overall results (f1 was higher by 3.5%), so we chose using it. Its also interesting to note that we gave higher weights to the positive labels (any relation is considered positive, while the 'NO_RELATION' label is considered negative) because there are much more negative samples then positive ones. This further improve the model's performance.

- **Features used:**
  - Entity type of the first chunk
  - Entity type of the second chunk
  - String of the first entity
  - String of the second entity
  - Indicator if entity 1 is a location

    - Indicator if entity 2 is a location
    - The word before the first entity
    - The word after the second entity
    - Entity of the head of the first entity in a dependency tree.
    - Entity of the head of the second entity in a dependency tree.
    - BOW between the two entities
    - POS tag list
    - word string list
    - dependency type list

- **Explanation regarding 2 interesting features:**
    - the 'is_location' features on the first and second chunk are important, because they provide some crucial extranal knowledge that is missing to the model. as expected, it helped increasing the model's performence. In order to create it, we have downloaded an external data source which contains a list of countries and cities.
    - the 'bag_of_words' features help the model to recognaize basic recurring patterns in the language. it contains the words between the two chunk in each sentence. although they are very noisy, because most of the times they won't contain sagnifiacnt patterns, the model was able to distingius their useful values. overall, it helped increasing the model's performance by a lot (~3.5%)

Error Analysis

Example of common recall errors

for example sentence 77:
Nikita Khrushchev instructed Soviet ships to ignore President Kennedy 's naval blockade during the Cuban missile crisis , but the order was reversed just hours before an inevitable confrontation , according to a new book

For this sentence our model predicted Kennedy live in Soviet which is a mistake.

The true answer is Nikita Khrushchev live in Soviet, where our model didn't relate to it (None)

The model needed external knowledge in order to predict keneddy's state. From reading the senetence it's unclear in what country kenedy lives. In order to exclude the option kenedy lives in soviet the model needed prior knowledge.

.

:Example of common precision errors

Those honored were Jean Griffith , mother of a man killed on a Howard Beach , Queens , parkway after being chased by a gang of whites in (December 1986 (sent 848

In this error we tagged “Live_In” where the answert was None.
The error was because Howard Beace was classified as country and not as a location in Queens.

In the sentence 862:
Soviet Foreign Eduard A. Shevardnadze is to visit China next month to pave the way for the first Chinese-Soviet summit in 30 years , Chinese television reported Monday .)

The model predicts the follwing prediction:
Eduara A. sheardnadze Live_In soveit
Eduara A. sheardnadze Live_In china
Eduara A. sheardnadze Live_In Chinese-Soviet

and the true answer is Eduara A. sheardnadze Live_In soveit.

Although, the model had a NER feature, it failed to conclude that a in every LIVE_IN realtion there muse be one word with NER = LOC.
It intreasting to mention that Rule-Based model would has dealt with it.

Results of the final model

	Percision	Recall	F1
Train set	0.766	0.827	0.795
Dev set	0.532	0.541	0.537