Submitters:

Ameen Ali	318429792
Tal Klein	312592520

Writeup 4 - Relation Extraction

Model Description:

We are focusing on the relation "**Live_In**". We have chosen to implement a hybrid system which combines the machine learning approach and the rule-based approach.

External resources:

- spacy
- scipy
- nltk (wordnet, punkt)
- numpy
- countries.json, cities.json, us.json
- sklearn – for a svc classifier

To implement the machine learning approach, we extracted features from the training file. The process of extraction is as follows:
First, for every sentence in the file, we extracted the named entities to chunks (extractChunks function) and then iterated over all possible chunk pairs in the given sentence. For each pair, we built its own features. As known, features are good indicators of entity relations and it is important to select those which are relevant to the "Live_In" relation.

The set of features we used are:

- entity of the head words
- the head words
- the head words' concatenation of the entities
- the POS tags of all the words in the chunk
- 2 words before and after each chunk (if exist)
- the words between the 2 chunks
- the pos tags of the words between the two chunks
- if the pos tag of the first word in the chunk start with NN (any type of noun), We added NN as a feature, otherwise, the POS tag of the word added.
- Using wordnet, all the synonyms of each word in the chuck were added as a feature
- the dependency tree (this feature represents the grammatical structure of a sentence)

- Using the json files of all possible countries, cities, and states in the U.S. - for each word, we checked if it was a country a city or a state, if yes, we added it as a feature

Note that we are only interested in chunks that have some relation between, therefore, we have extracted from the annotated train file the first part of the sentence and saved it in a map and for each chunk we encounter, we check if it in the map. In other words, we have linked between each chunk pair to its annotation in order to learn from this pair by their features.

After extracting all the mentioned features we have trained our model using the SVC classifier.

For the second approach, we have used a rule-based system as follows:

First, for each chunks, we check if both chunks contain a person and a location so they may have a relation of Live_In between them. If not, we continue to the next sentence.
If the prediction given us by the trained model is negative, we check if the first chunk entity is "PERSON" and the second chunk entity is "LOCATION" (using both isGazette function which checks whether the string is found in the countries / states / cities extracted from the json files and checkEntityChunkPhrase function which checks if a given entity is "PERSON")
If so, we check the rules between the two chunks in the following way:

- If the "person" chunk contains a word such as: "spokesman", "spokeswoman", "diver", "Lt.", "representative", "governor", "manager" which represent a word relation, before the chunk, we may assume that those jobs require the person to live in the place given.
- If between the chunks we encounter a word such as: "settle", "reside", "live", "living", "stay", "staying", "representative", this is an indicator of the place this person lives.
- If between the chunks we encounter a word such as "home of", "governor of", "king of" , "citizen of" , "resident of" we may assume that the person live in the given place.
- Chunks containing the word "manager" between them are not considered because they are probably an indicator of a "Word_For" relation.

Error Analysis:

Some of the errors in prediction were caused by a confusion with the Work_For / OrgBased_In tags because some of them contain a location. Learning on a small dataset is usually rather hard.

Most of the recall errors were caused by an incorrect tagging of a PERSON entity. We could improve it by having a dictionary of popular names or adding more rules.

Another improvement could be linking between organizations to locations, so if we know where the organization is located – we would know where the person lives. Another possible improvement could be checking adjectives that implies the nationality of the person – and that's probably where the person lives.

Precision errors (in prediction but not in gold)

Sentence 78: James G. Blight and David A. Welch of Harvard University 's John F. Kennedy School of Government say that ` ` if this order had held , war between the superpowers would probably have commenced at sea , shortly after 10 o 'clock on Wednesday morning , Oct. 24 , 1962 , several hundred miles off the coast of Cuba."

Gold: (Work_For!) James G. Blight Work_For Harvard University
James G. Blight Work_For John F. Kennedy School of Government
David A. Welch Work_For Harvard University
David A. Welch Work_For John F. Kennedy School of Government

Prediction: James G. Blight Live_In Cuba
David A. Welch Live_In Cuba

Analysis: the quotation contains a name of a place, this could be the source of the confusion, plus, a work_for tag confusion

Sentence 119: Composer Thomas is a native of Chisholm , Minn .

Gold: Thomas Live_In Chisholm
Thomas Live_In Minn .
Chisholm Located_In Minn .

Prediction: Composer Thomas Live_In Chisholm

Analysis: a. wrong tagging – composer got a PERSON tag
b. we did not have the abbreviation Minn. In our json files

Sentence 4660: James Earl Ray , 60 , is serving 99 years in prison for first-degree murder of King in Memphis on April 4 , 1968

Gold: James Earl Ray Kill King

Prediction: James Earl Ray Live_In Memphis

Analysis: confusion with the "Kill" tag

Recall errors (in gold but not in prediction)

Sentence 161: Britain 's Duchess of York and former Philippines first lady Imelda Marcos topped the list by Blackwell , a designer and self-appointed fashion arbiter whose real name is Richard Sylvan Selzer .

Gold: Duchess of York Live_In Britain

Prediction: -

Analysis: wrong tagging, Duchess of York tagged as ORG

Sentence 1049: The state Supreme Court jury in Niagara County made the award Tuesday to the estate of Thomas Viscomi and his widow , Norma Viscomi , of Niagara Falls .

Gold: Norma Viscomi Live_In Niagara Falls

Prediction: -

Analysis: wrong tagging, Niagara Falls tagged as a PERSON (?!?!?!)

Sentence 2887: Croatia 's Granic Admits Downed Planes Took Off From Croatia LD0303083794 Moscow ITAR-TASS in English 0729 GMT 3 Mar 94

Gold: Granic Live_In Croatia

Prediction: -

Analysis: wrong tagging, no PERSON tag for Granic

Conclusion:

Most of the mistakes were caused by wrong tagging.

Results:

Relation	Dev Recall	Dev Prec	Dev F1	Test Recall	Test Prec	Test F1
Live_In	0.25	0.37	0.3	0.58	0.77	0.66