Bar Ilan University
Natural Language Processing

Or Levitas

January 2023

This assignment is about Relation Extraction(RE) From a large corpus of texts. Two relations are requested:

1. Medical conditions and their treatments.
2. People who graduated from academic institutions.

The general approach for both of the task is the same:

A Golden relations - Get small data set with "golden" annotations. "Golden" annotations - means that the annotation is validated to be true.

B Model implementations - Train model on this data set.

C Model results - Examination of the model results.

D Additional relations - Train model on all the data. Then use it to achieve more relations by running it on a larger data set, obtained from variety of queries that may contain those relations.

1 Medical relations

1.1 Golden relations

The well known annotated data set published with the paper: "Classifying Semantic Relations in Bioscience Texts" paper by "Rosario et al."[2] is used as the golden data. The data set contain relations of diseases and their treatments from from the MEDLINE 2001 [1], annotated by UC Berkeley SIMS master student with a biological background. Example of such annotation in their corpus:

```
"CONCLUSION : <TREAT> Methylphenidate </TREAT> is effective in
treating children with <DIS> epilepsy </DIS> and <DIS> ADHD </DIS>
and safe in children who are seizure free"
```

From this data set a sample of 830 diseases and their treatments sentences and another 830 sentences without diseases or treatments were taken, for total balanced sample of 1660 sentences.

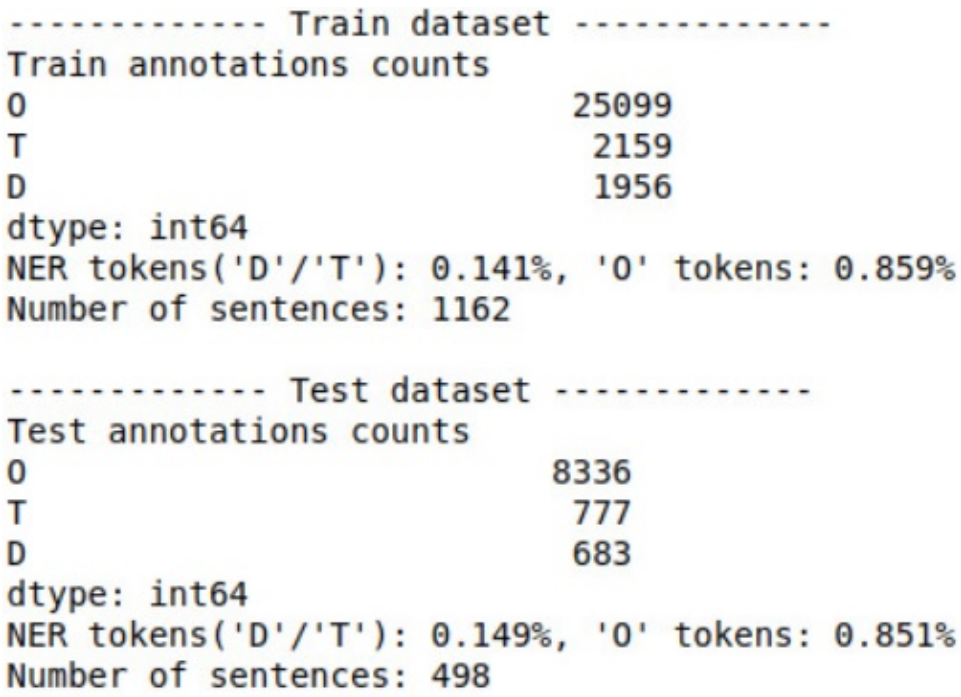

Figure 1: Statistics on the train and test data. The following are the annotations for each token: 'D' - disease, 'T' -treatment and 'O' - Other

This 1660 sample were splitted in a balanced way into, 70% train dataset (1162 samples) and 30% test dataset (498 samples). Fig. 1 show statistics on those datasets regarding entities annotations with 'D' - disease, 'T' -treatment and 'O' - Other.

1.2 Model implementation

Three CRF model were tested. The second the third are more as a baseline. The model try to extract the disease, the treatment entities and find the relation between (The relations could be learnt with the head and the dependent from dependency parsing).

The 3 types of models that have been learnt:

1. $Model_{Prev}$ - Previous and current words.
2. $Model_{Prev,Next}$ - Previous, current and next words.
3. $Model_{Range}$ - Range of 20 tokens. 10 to the left of the current token and 10 to the right. Fig. 2 shows an example of the feature with the relative location, the diseases are in the range of the treatment.

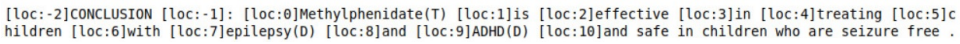

Figure 2: Example of the relative locations range taken from the left and right of the token "Methylphenidate". The entities annotations are in brackets ('D' - disease, 'T' -treatment and 'O' - Other). The diseases are in the range of the treatment

Each token have the following features:

- The word in lower case letters.
- The word Part of Speech(PoS).
- The last three letters of the word.
- The head from dependency parsing.
- The dependent from dependency parsing.
- The word in capital case letters.
- Indication whether the word is a number.
- The first letter of the words in capital case.
- "BEG" and "END" indications have been added to track begin and end of the sentences, respectively.

1.3 Model results

Model name	Accuracy	Precision	Recall	F1
$Model_{Prev}$	91.5%	86.3%	67.5%	74.4%
$Model_{Prev,Next}$	92.1%	88.1%	70.0%	77.0%
$Model_{Range}$	92.8%	88.5%	73.4%	79.5%

Although $Model_{Range}$ model is looking on a big range (20 tokens) it improvement from the linear CRF $Model_{Prev,Next}$ model in F1 is only 2.5%. Maybe because the head/dependent contain the enough information from a distance locations and the additional features from the distant tokens have little information (like Part of Speech). The best model is $Model_{Range}$ with F1 of 79.5% .

1.4 Additional relations

The Following queries were run with SPIKE on Pubmed data:

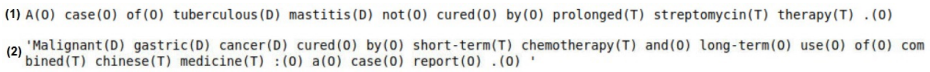

Figure 3: Example of two correctly relations extracted an query.

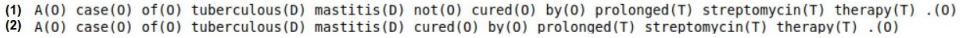

Figure 4: Example wrong extracted relation. Sentence (1) is an observed sentence in the corpus with correct relation. While the second, (2), is augmentation of the first with additional "not". The second sentence is false positive relation.

Spike Query	Records number
entity=DISEASE $cured $by tag=/NN.*/	2560
entity=DISEASE $cured $with tag=/NN.*/	4960
entity=DISEASE $treatment $with tag=/NN.*/	10000
entity=DISEASE $healed $by tag=/NN.*/	1932
entity=DISEASE $treated $by tag=/NN.*/	10000

In total 29452 records were extracted, out of them 29139 are unique. From those unique records 5013 relations were extracted, combined with the golden relations contain 1055 relations. So the total relations that were extracted are: **6068 relations**. Fig 3, shows examples of correctly extracted relations form the first query.

Another note, although cases like the one in Fig.4 have not been seen. It is hypothesised that those kind of sentences could be in the data. unfortunately, the model won't handle with those well.

2 Graduation relations

2.1 Golden relations

Since there are not known annotations, multiple queries on SPKIE, each manually annotated. Each query have the following pattern:

$$< E >: [entity]John \textbf{ pattern1 } < E >: [e = ORG]Yale \textbf{ pattern2}$$

The Queries for gold annotation are shown in Table 2.1. Those records were manually annotated for 200 positive samples and 24 negative samples.

```
-------------------- Datasets statistics: --------------------

'N' - Name
'I' - Institution
'O' - Other
------------- Train dataset -------------
Train annotations counts
O                           2546
I                            439
N                            186
dtype: int64
NER tokens('N'/'I'): 0.197%, 'O' tokens: 0.803%
Number of sentences: 156

------------- Test dataset -------------
Test annotations counts
O                           1195
I                            183
N                             89
dtype: int64
NER tokens('N'/'I'): 0.185%, 'O' tokens: 0.815%
Number of sentences: 68
```

Figure 5: Statistics on the train and test data. The following are the annotations for each token:'N' - Name of a person, 'I' -Institution and 'O' - Other.

pattern1	pattern2	records number
$was :[lemma]graduated $from	$university	31
$was :[lemma]graduated $from	$academy	4
$was :[lemma]graduated $from	$college	31
:[lemma]studied $in	$university	90
:[lemma]studied $in	$college	61
:[lemma]studied $in	$academy	18
:[lemma]educated $at	$university	4
:[lemma]educated $at	$college	15
:[lemma]educated $at	$academy	1
$did $not :[lemma]graduated $from		11
:[lemma]educated $at	$school	5
$was :[lemma]graduated $from	$school	21
:[lemma]studied $in	$school	114

Table 2.1. Queries for gold annotation.

These 224 samples were splitted in a balanced way into, 70% train dataset (156 samples) and 30% test dataset (68 samples). Fig. 5 show statistics on those datasets regarding entities annotations with 'N' - Name of a person, 'I' -Institution and 'O' - Other.

Multiple assumptions regarding the manually annotated data (and consequently what the model will learn) were made by examine the dataset. For each assumption a brief explanation with an example that illustrate the spirit of the assumption are shown. The assumption are:

1. *"Oliver Wendell Holmes , Jr. was graduated from Harvard Law School in 1866 , and opened a private law practice ."*. In the example, we can see that *school* can be also an academic institute, so school not automatically classifier as negative samples.

2. *"Kanodia studied science in Elphinstone College , Bombay University for 2 years ."*. Since the number of year is specifically mentioned, it can be deduced that it is irregular to learn for 2 year and therefore, classified as wrong sample. In other words, specification meant irregular and irregularity assumed to mean failing to graduate.

3. *"Ho studied Philosophy in Lingnan University ."*. An assumption that if not mentioned otherwise, it is presumed that a person that studied in an academic institute, also finished.

4. *"Chan studied in CCC Heep Woh College from 1972 to 1976"*. Although it is specified, a range usually meant that a person finished his studies (by looking at Wikipedia pages).

5. *"Before the war , Korngold studied journalism in the Warsaw School of Political Sciences ."*. If interruption in mentioned, that person did not finished his studies.

6. *"Donald graduated from London School of Art"*. Unclear place of graduation (is it a school or academic institute) were skipped. This is why the total queries records is not equal to the number of gold annotations.

2.2 Model implementation

As in the medical data. three CRF model were tested. However, because I have more knowledge on graduation than on medical knowledge (For example, common names of academic institutes), those four additional features were add to those shown in the previous section.

- None academic school indication - Does the name in the list: ['high', 'primary', 'intermediate', 'secondary']..
- Academic indication- Does the name in the list: ["university", "college", "institute", "academy", "institution"].
- Person Named Entity Recognition indication - Does the word is in the range of the PERSON NER of the sentence.
- Organisation Named Entity Recognition indication - Does the word is in the range of the ORG NER of the sentence.

2.3 Model results

Model name	Accuracy	Precision	Recall	F1
$Model_{Prev}$	94.3%	90.0%	90.5%	89.3%
$Model_{Prev,Next}$	95.5%	91.4%	91.8%	91.4%
$Model_{Range}$	96.5%	93.0%	92.3%	92.5%

Although $Model_{Range}$ model is looking on a big range (20 tokens) it improvement from the linear CRF $Model_{Prev,Next}$ model in F1 is only 0.9%. Although additional features were introduced and although the NER should be a strong features. The higher F1 score of 92.5% in graduation relative to the medical f1 score of 79.5% could also be explained by either relative smaller datasets or easier domain (medical is harder domain for RE than graduation).

2.4 Additional relations

The Following queries were run with SPIKE on Wikipedia data:

Spike Query	Records number
entity=PERSON graduated at entity=ORG	14140
entity=PERSON graduated from entity=ORG	45924
entity=PERSON studied in entity=ORG	45065

In total 105353 records were extracted, out of them 94022 are unique. From those unique records 29114 relations were extracted, combined with the positive golden relations contain 200 relations. So the total relations that were extracted are: **29314 relations**.

References

[1] *PubMed Data set*. URL: `https://www.nlm.nih.gov/databases/download/pubmed_medline.html`.

[2] Barbara Rosario and Marti A. Hearst. "Classifying Semantic Relations in Bioscience Texts". In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*. ACL, 2004, pp. 430–437.