Course : CSC591/791, Graph Data Mining
Project: Named-Entity Recognition
Unity-ID: akagrawa

1 Project Description

Named-Entity Recognition (NER) is a task in the information extraction pipeline that seeks to locate and classify different elements present in a text into predefined categories, such as the names of persons, organizations, locations, date, time etc. It is one of the key steps in building Semantic Webs and constructing semantically aware search engines. Typically, NER tasks are accomplished by using either rule based methods or probabilistic methods. Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) are probabilistic graphical models commonly used to build NER models. CRF models usually offer a better performance than HMMs because they can take into consideration both past and future states. These models are typically trained using supervised learning methods, i.e., trained using data with known entity tags[2].

2 Performance Evaluation

DataSet1 : Wikipedia
Model : ner.wiki.model1.ser.gz
Parameters: Please refer README.md file for the detail description of the parameters tuning for training the models. The table1 is the performance metrics for default parameters.

Entity	LOC	MISC	ORG	PERS	TOTAL
Precision	0.6336	0.5851	0.4415	0.8079	0.6874
Balanced Accuracy	0.7603	0.6669	0.7651	0.8239	0.7836
Recall	0.5319	0.3437	0.5492	0.6549	0.6149
F1-Score	0.5783	0.4330	0.4895	0.7231	0.6412
True Positive Rate	0.5319	0.3437	0.5492	0.6549	0.6149
True Negative Rate	0.9827	0.974	0.9876	0.9856	0.9789
False Positive Rate	0.0173	0.0260	0.0123	0.0144	0.0211
False Negative Rate	0.468	0.6563	0.4508	0.3456	0.3851

Table 1: Performance Measure for Dataset1: Wikipedia Test

DataSet2 : Emma
Model : ner.emma.model1.ser.gz
Parameters: Please refer README.md file for the detail description of the parameters tuning for training the models. The table2 is the performance metrics for default parameters.

Entity	LOC	PERS	TOTAL
Precision	0.8181	0.9210	0.9091
Balanced Accuracy	0.7994	0.8959	0.8657
Recall	0.6000	0.7954	0.7977
F1-Score	0.6923	0.8537	0.8463
True Positive Rate	0.6000	0.7954	0.7977
True Negative Rate	0.9966	0.9894	0.9800
False Positive Rate	0.003	0.0106	0.0199
False Negative Rate	0.4	0.2045	0.2023

Table 2: Performance Measure for Dataset2: Emma Test

3 Questions

1. **For which of the labels did you obtain the best performance? Is this performance due to having more training data for that particular label?**

   Solution : The performance with reference to the above performance table follows the pattern below. In the training data, there are more samples of PERS than that of ORG and LOC etc. Hence having more training data for the particular label has certainly infludence the performance.

   performance(PERS) > performance(LOC) > performance(ORG) > performance(MISC)

   Note: In Emma Dataset the ORG and MISC labels are not present in train and test data.

2. **Do the models perform better during training for the Wikipedia (Dataset 1) data or for the Emma (Dataset 2) data? Speculate why.**

   Solution : In order to compare the training performance of both the dataset, the performance measures on training dataset iteself is needed to be calculated.

For Wikipedia (Dataset1) : The performance output on the training data:

Entity	P	R	F1	TP	FP	FN
LOC	0.9990	0.9941	0.9965	1008	1	6
MISC	0.9986	0.9930	0.9958	707	1	5
ORG	0.9955	0.9955	0.9955	894	4	4
PERS	0.9989	0.9968	0.9979	931	1	3
Totals	0.9980	0.9949	0.9965	3540	7	18

Table 3: Performance Measure for Dataset1: Wikipedia Train

For Emma (Dataset2) : The performance output on the training data:

Entity	P	R	F1	TP	FP	FN
LOC	1	1	1	14	0	0
PERS	1	1	1	98	0	0
Totals	1	1	1	112	0	0

Table 4: Performance Measure for Dataset1: Emma Train

From the above performance measures, we can conclude that the model perform better during the training for the Emma(Dataset2) as compare to the training for the Wikipedia(Dataset1). The above observation can be due to the small training data size of the Emma dataset as compare to Wikipedia dataset. Also in the Emma Dataset the variability of the words were less i.e. there were very few unique words(Mr, Mrs, Emma, Taylor, London etc) which where tagged as entities as compare to Wikipedia set.

3. **Do the models perform better during testing for the Wikipedia (Dataset 1) data or for the Emma (Dataset2) data? Speculate why.**

Solution : In order to compare the test performance of both the dataset, the performance measures on test dataset is needed to be calculated. We can compare the performance measures listed in table 1 and table 2.
Based on the observations, we can say that again the performance measures for Emma Dataset2 is better than that of Wikipedia Dataset1. The reasons for the observations can be analyzed as below:

- The training performance of both the datasets, the testing performance can also be extrapolated in this case.
- The test data sizes have a massive size difference. The testdata size of wikipedia is much more than its training data. Whereas the testdata size of emma is less than its training data. This means that the wikipedia training data might not have enough representative samples.
- Since the emma testdata has the almost the similar words as observed in the training data, the test accuracy is high which is comparatively less in the wikipedia dataset.

4. **"Research indicates that even state-of-the-art NER systems are brittle, meaning that NER systems developed for one domain do not typically perform well on other domains." Is this true for your CRF models? In other words, do the models trained using Wikipedia (Dataset 1) or Emma (Dataset 2) data, work well for Twitter (Dataset 3) data?**

Solution: For the performance measurements of both CRFmodels on twitter testdata, we need to compare the performance output on twitter data. The tables below lists the performance metrics. Using CFRmodel of Wikipedia(Dataset1) : The performance output on the twitter test data:

Entity	P	R	F1	TP	FP	FN
LOC	0.4725	0.3431	0.3975	129	144	247
MISC	0.1209	0.1162	0.1185	51	371	388
ORG	0.0258	0.0468	0.0333	8	302	163
PERS	0.5301	0.2803	0.3667	141	125	362
Totals	0.2589	0.2210	0.2384	329	942	1160

Table 5: Performance Measure for Twitter Dataset using Wikipedia CRFmodel

Using CFRmodel of Emma(Dataset2) : The performance output on the twitter test data:

Entity	P	R	F1	TP	FP	FN
LOC	0.3480	0.1888	0.2448	71	133	305
PERS	0.1779	0.2147	0.1946	108	499	395
Totals	0.2207	0.1202	0.1557	179	632	1310

Table 6: Performance Measure for Twitter Dataset using Emma CRFmodel

Based on the above observations we can say that these models have not performed better for the twitter test data. The precision and recall for individual labels is within 15-20% in average which is very less. The CRFModel build on twitter training data must have better performance as compare to the above results.

5. **Does the distribution of each label in the train and test data affect how well a particular label is classified?**

   Solution: Based on the observations of the performance measures on the train and test data, we can conclude that the label distribution has definitely an influence on the prediction accuracy and related measures. We see that,

   performance(PERS) > performance(LOC) > performance(ORG) > performance(MISC)

   And number of labels also follows somewhat the same pattern,
   count(PERS) > count(LOC) > count(ORG) > count(MISC)

   Data with proper distribution which captures maximum variability in the data, tend to produce better models when applied with an appropriate classifiers.

6. **With the amount of data available and the number of different possible labels that can be assigned, a manually annotated corpus to train an NER model might be difficult to find. Moreover, when the model is not trained with good quality data, the annotations assigned by the model are often incorrect. What are some measures that can be put in place to train better NER models? Please, let your imagination run "wild" but keep in mind what can be accomplished computationally and what cannot.**

   Solution: In order to train better models the quality of the training data labels must be significantly high. In the scenario of Named-Entity Recognition a manually corpus is a difficult task to achieve.
   Hence the automation in this process can significantly provide more opportunities to

develop better and more efficient classifiers for this domain.
We can follow certain heuristics for the automated labeling of the corpus.

- Pattern recognition for certain attributes, like Location, always starts with capital letter, followed by lowercase alphabets. "Xxxxx" Similar patterns can be applied for some common labels.
- Understanding the structure of the sentence. For example: Sentences are primarily subdivided into Subject + Predicates. If the model is able to detect subjects and objects in the sentence, certain Name labels can be labeled with more ease.

7. **Additional questions for the Emma (Dataset 2) data:**

   a. **Which set of 2 labels took the longest time to train? Is this time positively or negatively correlated with the amount of data available for that label set?**

Labels	Count of Distinct Labels	Training Time in(Secs)
O,PERS	3187, 160	2.5 sec
O,LOC	3331, 16	2.7 sec
O,ORG	3345, 2	2.5 sec

Table 7: Training Time for CFRModel on Emma Dataset

From the above observation, we can see that the label **{O, LOC}** took the longest time(2.7 secs) to train. With reference to the training time, we can see no correlation between the amount of data available and time taken in this case. But this case cannot be generalized as the sample size is too less. Note: Please find different datasets for the above combination, inside folder data.

   b. **Which set of 3 labels took the longest time to train? Is this time positively or negatively correlated?**

Labels	Count of Distinct Labels	Training Time in(Secs)
O,PERS,LOC	3171, 160, 16	3.5 sec
O,ORG,LOC	3329, 2, 16	3.7 sec
O,PERS,ORG	3185, 160, 2	4.0 sec

Table 8: Training Time for CFRModel on Emma Dataset

From the above observation, we can see that the label **{O, PERS, ORG}** took the longest time(4.0 secs) to train. With reference to the training time, we can see no correlation between the amount of data available and time taken in this case. But this case cannot be generalized as the sample size is too less.

c. **Which set of 2 labels produced the best test results? Are these results positively or negatively correlated with the training performance of that label set?**

Here the performance of label "O" is avoided. And based on the other label performance, the comparison is made.

Entity	Total(P)	Total(R)	Total(F1)	Total(TP)	Total(FP)	Total(FN)
O, PERS	0.8056	0.5800	0.6744	29	7	21
O, LOC	0.8750	0.4667	0.6087	7	1	8
O, ORG	0	0	NA	0	0	3

Table 9: Performance Measure for Emma Test Dataset for 2 Labels

From the above observation, we can see that the label **{O, PERS}** have better F1-score. This clearly reveals that there exist a positive correlation between the performance measures (F1-score, Recall) and the label set size in this case. But this case cannot be generalized as the sample size is too less.

d. **Which set of 3 labels produced the best test results? Are these results positively or negatively correlated with the training performance of that label set?**

Entity	Total(P)	Total(R)	Total(F1)	Total(TP)	Total(FP)	Total(FN)
O, PERS, LOC	0.7273	0.6154	0.6667	40	15	25
O, ORG, LOC	0.8750	0.3889	0.5385	7	1	11
O, PERS, ORG	0.7838	0.5472	0.6444	29	8	24

Table 10: Performance Measure for Emma Test Dataset for 3 Labels

From the above observation, we can see that the label **{O, PERS, LOC}** have better F1-score. This clearly reveals that there exist a positive correlation between the performance measures (F1-score, Recall) and the label set size in this case. But this case cannot be generalized as the sample size is too less.

e. **Randomly assign labels to the data, as follows:**

i. **For the set of 4 labels (O, PERS, ORG, LOC), assign 0 to O, 1 to PERS, 2 to ORG, and 3 to LOC. For each word in the training and test datasets, calculate the line number of the word mod 4 and assign the corresponding label to the word. What is the performance of the model built using these labels? Does any label perform better than the model built using the manually curated data?**

Entity	P	R	F1	TP	FP	FN
O [0]	0.1869	0.1865	0.1867	83	361	362
PERS [1]	0.1888	0.1888	0.1888	84	361	361
ORG [2]	0.1865	0.1865	0.1865	83	362	362
LOC [3]	0.1869	0.1865	0.1867	83	361	362
Totals	0.1873	0.1871	0.1872	333	1445	1447

Table 11: Performance Measure for Emma using Random Labels for 4 Labels

Based on the above observations, it is clear that randomly assigning data has resulted in very poor performance as compare to the performance results obtained from manually labeled data.

ii. **For each set of 3 labels (e.g., O, PERS, LOC), assign 0 to O, 1 to PERS, and 2 to LOC. For each word in the training and test datasets, calculate the line number of the word mod 3 and assign the corresponding label to the word. What is the performance of the model built using these labels? Does any label or label set perform better than the model built using the manually curated data?**

Entity	P	R	F1	TP	FP	FN
O [0]	0.1216	0.1214	0.1215	72	520	521
PERS [1]	0.1214	0.1212	0.1213	72	521	522
LOC [2]	0.1214	0.1214	0.1214	72	521	521
Totals	0.1215	0.1213	0.1214	216	1562	1564

Table 12: Performance Measure for Emma using Random Labels for 3 Labels

Based on the above observations, it is clear that randomly assigning data has resulted in very poor performance as compare to the performance results obtained from manually labeled data. The performance results are even bad than that of assigning 4 random labels as shown in table 11 and table 12.

4 Parameter Tuning

Below are the aggreated performance results(Total) on different parameters of CRF Model, for Wikipedia and Emma Dataset. Only changed parameter from the default set is listed in the tables.

For WikiPedia TestData Performance

Parameter	**P**	**R**	**F1**	**TP**	**FP**	**FN**
Default	0.5799	0.4909	0.5317	150109	108745	155673
Maxleft=2	0.5798	0.4897	0.5309	149734	108531	156048
Maxleft=3*	NA	NA	NA	NA	NA	NA
maxNGramLeng=4	0.5733	0.4893	0.5280	149619	111338	156163
useNgram=false	0.1216	0.1214	0.1215	72	520	521

Table 13: Performance Measures for different set of parameters for Wikipedia Dataset

*System out of memory.

For Emma TestData Performance

Parameter	**P**	**R**	**F1**	**TP**	**FP**	**FN**
Default	0.7818	0.6324	0.6992	43	12	25
Maxleft=2	0.7679	0.6324	0.6935	43	13	25
Maxleft=3**	0.7679	0.6324	0.6935	43	13	25
maxNGramLeng=4	0.8000	0.6471	0.7154	44	11	24
useNgram=false	0.7143	0.5882	0.6452	40	16	28

Table 14: Performance Measures for different set of parameters for Emma Dataset

Inference

- Performance is low, if we do not include Ngram i.e. useNgram=false.
- If the maximum left and right words are varied in model building, the more left or right words are considered the more time and memory for model builiding is necessary.
- MaxNGramLength = 4 gives almost the same(slightly less) performance measures as compare to MaxNGramLength = 6 (Default Parameter).

5 References :

[1] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/ manning/papers/gibbscrf3.pdf

[2] Kanchana Padmanabhan, Project Description Document- Named Entity Recognition.

[3] Nathalie Japkowicz, Mohak Shah, Performance Evaluation for Learning Algorithms: Techniques, Applications, and Issues, Tutorial at International Conference on Machine Learning, 2012, Edinburgh, Scotland.