Machine Learning for Named Entity Recognition Report

Melis Nur Verir

Abstract

This article focuses on gaining a thorough knowledge of the problem and making sure the appropriate materials are available for designing the system, as well as the first preparations for building and assessing systems for Named Entity Recognition. The background information about the named entity identification task's exploration is given, and data exploration techniques along with the fundamental setup for a typical assessment are investigated.

1 Introduction

First, the necessary information about the task and named entity recognition will be mentioned in Section 2. Then, the necessary information about the data set used in this study will be provided. Then, in Section 3, the models used in the study and the findings obtained as a result of feature discovery will be discussed. Finally, the results and evaluation of the experiments will be included in Section 4.

2 Task and Data

2.1 Task

Recognition of named entities inside text is the goal of the task of named entity recognition (NER), which belongs to the domain of natural language processing. NER systems are frequently implemented as the initial stage of information retrieval, topic modeling, co-reference resolution, and question answering(Yadav and Bethard, 2019). Language-independent named entity recognition is the focus of CoNLL-2003's joint task, and according to (Sang and De Meulder, 2003) integrating the unannotated data into the learning process was the difficulty of this task. They concentrated on four categories of named entities: people, places, organizations, and names of other entities not included in the first three categories.

2.2 Dataset and Distribution

The experiments include a training file, a development file, a test file, and a sizable file with unannotated data for each language, and the training data is used to train the learning algorithms. The English data is provided from Reuters news articles from the months of August 1996 and August 1997. Ten days' worth of data from the files, indicating the end of August 1996, was taken for the training and development set, while the texts for the test set were dated December 1996. According to (Ratinov and Roth, 2009) the test dataset is hence far more challenging than the development set since the named entities that are referenced in the test dataset differ significantly from those in the training set or development set.

The CoNLL-2003 named entity data includes one word per line, with empty lines serving as the beginning and end of sentences. There is a tag that indicates whether the current word is inside a named entity or not at the end of each line, and the type of the specified entity is also included in the tag. Entities of four different categories are contained in the data: persons (PER), organizations (ORG), locations (LOC), and miscellaneous names (MISC). The tagging system is the original IOB system, which was proposed by (Ramshaw and Marcus, 1995). When a named entity is embedded in another named entity, typically only the top-level entity has been annotated. The study makes the assumption that named entities are non-recursive and non-overlapping.

The Table 1 is a representation of the train data from CoNLL-2003 named entity data. When the second, third, and fourth columns of the data are examined, it is seen that these columns contain chunks. The words that makeup chunks are classified using part-of-speech tags, which allow one to specify a pattern or exclude certain words from the chunk. IOB tags are a format for chunks. Similar to part-of-speech tags, these tags can indicate the

EU	NNP	B-NP	B-ORG
rejects	VBZ	B-VP	O
German	JJ	B-NP	B-MISC
call	NN	I-NP	O
to	TO	B-VP	O
boycott	VB	I-VP	O
British	JJ	B-NP	B-MISC
lamb	NN	I-NP	O
.	.	O	O

Table 1: Example of the CoNLL-2003 named entity data

inside, outside, and commencement of a chunk in addition to parts of speech. Every word has a part-of-speech tag, and an IOB tag, and is on its own line. B-NP denotes the start of a noun phrase, I-NP shows that the word is contained inside the noun phrase, and O indicates the ending of the sentence. B-VP and I-VP represent the start and middle of a verb phrase, respectively.

2.3 Preprocessing

Data preprocessing includes linguistic preprocessing of raw data such as a tokenizer, part-of-speech tagger, and chunker, and is made available after these processes are applied (Sang and De Meulder, 2003). In this analysis, there was no additional preprocessing done to the data.

2.4 Evaluation Metrics

In the evaluation part, mini datas were used. First, gold annotation and machine annotation were extracted from the given data and the extracted gold annotations were compared with the machine output and counted. Confusion matrix tables were created with these comparison results. The confusion matrix is an approach used to see how well the prediction model is performing. The number of predictions the model makes when it correctly or incorrectly classifies classes is indicated by each entry in a confusion matrix. The number of predictions when the classifier properly identifies the positive class as positive is known as True Positive (TP). The number of predictions in which the classifier properly identified the negative class as negative is known as True Negative (TN). False Positives (FP) are instances in which a classifier predicts a negative class as a positive one. False Negative (FN) term describes the proportion of predictions in which the classifier misinterprets the positive class as the negative class.

Then, precision, recall and fscore were calculated for each class with these metric values. Measures of software reliability include precision, recall, and F-score, which provide data about an analyzed document. Precision is a percentage statistic that indicates how many of a set of outcomes from a processed document are accurate.

$$Precision = \frac{TP}{TP+FP} \; Recall = \frac{TP}{TP+FN}$$

Recall is a percentage metric that represents the proportion of accurate findings discovered. The harmonic mean of a system's Precision and Recall values is known as an F-score, and it can be calculated using the following formula:

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN}$$

3 Models and Features

3.1 Logistic Regression

Logistic Regression which is a predictive analytic algorithm based on the idea of probability, logistic regression is a machine learning approach that is used for categorization problems and was used as a model in this study. The logistic sigmoid function is used by logistic regression to convert its output and produce a probability value (Nick and Campbell, 2007).

First of all, the properties and labels of the given data were extracted by assigning data and targets, and the Logistic Regression classifier was created. Since the size of the data in this study is enormous, the maximum iteration parameter was assigned as 3000. DictVectorizer is used for feature extraction, which converts feature-value mapping lists to vectors. The model was created by attaching the data used for training to this vectorizer. Another feature extraction method, TfidfVectorizer, which combines the features of TfidfTransformer and CountVectorizer into a single model, was used to better evaluate the results.

3.2 Support Vector Machine

The Support Vector Machine (SVM) is a classifier that separates positive and negative samples along a solid line in the middle known as the decision line in order to determine the best hyperplane between two classes of data. The SVM classifier folds all other classes simultaneously from a single model for all classes (Guia et al., 2019). The following two factors put SVMs ahead of more traditional statistical learning algorithms like Decision Trees, Hidden Markov Models, and Maximum Entropy Models: SVMs perform well in terms of generalization regardless of the size of the feature vectors. By including the Kernel function, SVMs may perform their learning with all possible combinations of the input characteristics without raising the computational cost (Ekbal and Bandyopadhyay, 2010).

A NER system developed by (Asahara and Matsumoto, 2003) based on Support Vector Machines (SVMs) for Japanese speakers, and his method is an expansion of Kudo's chunking system, which performed the best on CoNLL-2000 shared tasks.

3.3 Features

The distribution of NER tags was examined to further examine the data. While the training data set was 203621, the validation data was found to be 51362. When both data sets of NER labels are examined, it is seen that the sample of the O label is in the first place with the highest number. However, since validation data has a greater size than train data, the distribution of the O tag in train data is lower. While this number is 169578 for the training data set, it is 42759 for the validation set. Table 2 is shown when the number of labels is proportional to the length of the data. Accordingly, the tags with the least number of examples in the training data set are I-MISC and I-LOC tags.

	count
O	169578
B-LOC	7140
B-PER	6600
B-ORG	6321
I-PER	4528
I-ORG	3704
B-MISC	3438
I-LOC	1157
I-MISC	1155

Table 2: the distribution of the NER labels in train data

The first 50000 data of the training dataset were extracted and analyzed as a sample to examine the most common words. As a result of the analysis, it was determined that the word *Thursday* was repeated 141 times and was the most used word. This is followed by the word *1996-08-22*, which repeats 125 times. Examining this result, it can be seen that at the time the word represents, Bill Clinton signed the Personal Responsibility and Work Opportunity Act, also known as *welfare reform*, which cut federal aid to vulnerable citizens. The reason is, of course, that the dataset was taken from Reuters news articles in August 1996 and August 1997, with data in English.When companies and organizations are sorted out using the ORG tag, it can be seen from the Figure 1 that Newsroom, Reuters, and Lauck appear in the top three.

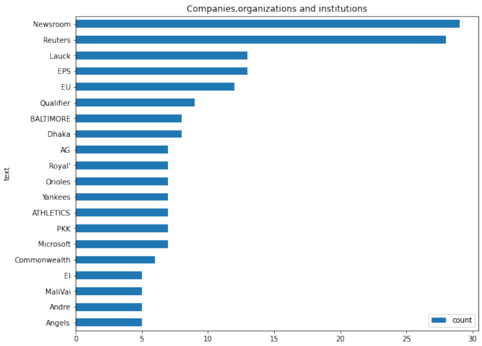


Figure 1: Companies,organizations and institutions

3.4 Feature Extraction

When the studies were examined, the features used in most studies were determined and they were applied to the named entity data of CoNLL-2003. The applied properties are lemmatization, stemming, previous-next token, and whether the token is capitalized, respectively. *Stemming* refers to a primitive heuristic method that removes derivational affixes from words as part of its attempt to achieve this aim most of the time. *Lemmatization* is the process of appropriately using a vocabulary and morphological analysis of words with the goal of removing only inflectional ends and returning the lemma, or dictionary form, of a word. Proper *capitalization* has been implemented as it makes proper names (organization, company, institute names) easier to detect and helps improve accuracy in classification.

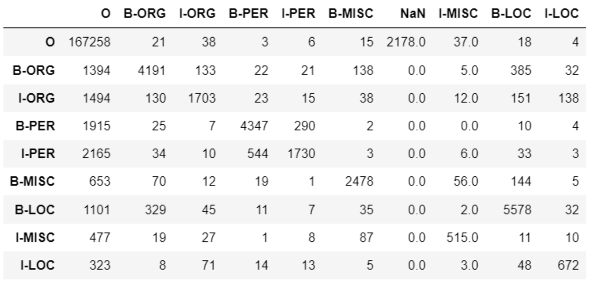
token	**pos**	**tag**	**ner**	**previous**	**latter**	**capitals**	**stemm**	**lemma**
rejects	VBZ	B-VP	O	EU	German	reject	reject	reject
German	JJ	B-NP	B-MISC	rejects	call	0	german	German

Table 3: An example of extracted features of train entity data of CoNLL-2003

4 Experiments and Results

4.1 Evaluation

In Figure 3, labels were counted and a confusion matrix was provided for each class. For example, when the *O* label with the most distribution in the data set was examined, it was seen that it was predicted correctly 167258 times. While the tag was actually *B-ORG*, it was estimated 1394 times as the *O* tag. The NaN value in the table in Figure 3 represents blank lines that serve as the beginning and end of sentences. No preprocessing has been done for this in the provided data set, but this value can be extracted in future studies for a more reliable result.



	O	B-ORG	I-ORG	B-PER	I-PER	B-MISC	NaN	I-MISC	B-LOC	I-LOC
O	167258	21	38	3	6	15	2178.0	37.0	18	4
B-ORG	1394	4191	133	22	21	138	0.0	5.0	385	32
I-ORG	1494	130	1703	23	15	38	0.0	12.0	151	138
B-PER	1915	25	7	4347	290	2	0.0	0.0	10	4
I-PER	2165	34	10	544	1730	3	0.0	6.0	33	3
B-MISC	653	70	12	19	1	2478	0.0	56.0	144	5
B-LOC	1101	329	45	11	7	35	0.0	2.0	5578	32
I-MISC	477	19	27	1	8	87	0.0	515.0	11	10
I-LOC	323	8	71	14	13	5	0.0	3.0	48	672

Figure 2: Companies,organizations and institutions

Table 4 shows the results of the evaluation metrics. When the Precision section is examined, it is seen that 95% of all labels that the model predicts as *O* actually have *O* labels. Looking at the *I-LOC* label, it is seen that this rate is very low. The main reason for this is that the label distributions examined in the 3.3 section are not equal.

	Precision	**Recall**	**F1-Score**
O	0.9461	0.9863	0.9658
B-ORG	0.8682	0.6630	0.7518
I-ORG	0.8323	0.4597	0.5923
B-PER	0.8721	0.6586	0.7505
B-MISC	0.8273	0.3820	0.5227
I-MISC	0.884	0.7207	0.7943
B-LOC	0.8097	0.4458	0.5750
I-LOC	0.0075	0.0414	0.0127

Table 4: Results of train data's precision recall and fscore for each class

Of all the words that actually had the O tag, the model predicted this result correctly only for 98% of those words. Likewise, the *I-MISC* tag is accepted as correct with 72%. The same performance cannot be reached in *I-LOC*. This may be because place names conflict with person names and organization, and the *I-LOC* tag is less in number compared to other tags. If the F1 value is not close to 1, it indicates that the model is doing a poor job of predicting the labels. Based on this finding, it can be seen that the model does a good job in predicting O, I-MISC, B-PER and B-ORG tags, but performs poorly on other tags.

References

Masayuki Asahara and Yuji Matsumoto. 2003. Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 8–15.

Asif Ekbal and Sivaji Bandyopadhyay. 2010. Named entity recognition using support vector machine: A language independent approach. *International Journal of Electrical and Computer Engineering*, 4(3):589–604.

Márcio Guia, Rodrigo Silva, and Jorge Bernardino. 2019. Comparison of naïve bayes, support vector machine, decision trees and random forest on sentiment analysis. pages 525–531.

Todd G Nick and Kathleen M Campbell. 2007. Logistic regression. *Topics in biostatistics*, pages 273–301.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

A Time spent

Please use Table 5 to give an overview of the time you spent on each submission.

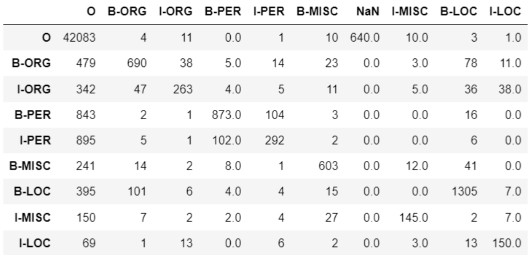
Week	Task	Time
1 1	lecture videos research	1 hour 3 hours
2 2	coding report	6 hours 4 hours

3
3
	coding(feature extraction)
	report(theory)
		2 hours
		2 hours

Total	18hours	

Table 5: Time overview.

B Appendix



	O	B-ORG	I-ORG	B-PER	I-PER	B-MISC	NaN	I-MISC	B-LOC	I-LOC
O	42083	4	11	0.0	1	10	640.0	10.0	3	1.0
B-ORG	479	690	38	5.0	14	23	0.0	3.0	78	11.0
I-ORG	342	47	263	4.0	5	11	0.0	5.0	36	38.0
B-PER	843	2	1	873.0	104	3	0.0	0.0	16	0.0
I-PER	895	5	1	102.0	292	2	0.0	0.0	6	0.0
B-MISC	241	14	2	8.0	1	603	0.0	12.0	41	0.0
B-LOC	395	101	6	4.0	4	15	0.0	0.0	1305	7.0
I-MISC	150	7	2	2.0	4	27	0.0	145.0	2	7.0
I-LOC	69	1	13	0.0	6	2	0.0	3.0	13	150.0

Figure 3: The confusion matrix of validation data