Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 4 No. 1, 2019, pp. 1 – 8

A study on Named Entity Recognition with OpenNLP at English Texts

Metin BİLGİN*[1]
Department of Computer Engineering, Bursa Uludağ University, Turkey
*E-mail: metinbilgin@uludag.edu.tr**[1]*
**Corresponding author*

Abstract - Named entity recognition is a subject, inside of information retrieval which is a subdomain of natural processing. It pertains to identifying and labeling of location, person, organization, etc., inside of text content. Named entity recognition provides identifying and classifying of person, area, etc. inside of formal and informal text content and it can be used for different purposes as question answering systems and removal of the relation between events. In this work, named entity recognition is performed and one method is suggested, and results are discussed for assignment to unlabeled name entities by using OpenNLP library with the help of KNIME program in the data set.

Keywords - Deep Learning, Named Entity Recognition, Natural Language Processing, OpenNLP, Text Mining

1. INTRODUCTION

Named entity recognition (NER) is a subdivision of information retrieval which is a part of the working area of natural language processing, and it is about identification and classifying [1].

Purpose of NER is to find entity names and classifying them as a person, location, area, time, a monetary statement which is dependent or independent to language in all documents which are dependent to a working area or independent to the working area and formal or informal.

NER's purpose is to classify entity names in the text with labeling them with classifies which are determined before processing. Fig. 1 includes an example of a text which is labeled for named entity recognition.

Figure 1. Named Entity Recognition [2].

As seen in Fig. 1, person, organization and location names are labeled, time money and date names are also used with this commonly used names

NER systems can find an area in many application of Natural Language Processing (NLP). NER is a subtopic of information retrieval and is used as an important stage for reaching information, besides it has been benefiting in search engine, multimedia indexing, machine translation, sentiment analysis, NLP application [3].

NER has two approaches which are rule-based and statistical, on the other hand, there are hybrid systems which are using these two approaches together.

In rule-based approaching, NER is being made by using grammar rules, while statistical approaching, is using a statistical model which is trained by using machine learning techniques. The first systems have been using ruled based algorithms which are defined manual, on the other hand, next systems are commonly using ordered labeling algorithm or automatic rule extraction with machine learning techniques [1].

One of the first works in this area is the identification of corporate names in the text which is made by Rau[4], studies in this area are being evaluated with language, document type and area and type of entity names [1]. Chiu and Nichols have presented a novel neural network architecture and using a hybrid bidirectional long-short-term-memory (LSTM and Convolutional neural networks (CNN) architecture for NER [5]. Derczynski et al. are described as a new Twitter entity disambiguation dataset, and conduct an empirical analysis of named entity recognition and disambiguation [6]. Santos and Guimaraes are based on the CharWNN deep neural network, which uses word-level and character-level representations (embeddings) to perform sequential classification for NER [7]. Marrero et. al. are analyzed the evolution of the NER field from a theoretical and practical point of view [8]. Shaalan is attempted to describe and detail the recent increase in interest and progress made in Arabic NER research and the different Arabic linguistic resources are presented and the approaches used in Arabic NER field are explained [9].

Learning methods are divided into three groups which are being used in name entity recognition. These groups are fully trained, semi-supervised and without training. Hidden Markov Models [10], Decision Trees [11], Maximum Entropy Markov Model [12], Support Vector Machines [13] and Conditional Random Fields [14-15] can be an example for fully trained systems.

In these systems, a marked wide corpus, tangible assets list, and rules to resolve uncertainty are being used. In the semi-supervised system, in the beginning, we use a cluster for training after that system learns by itself [16]. In unsupervised learning systems, we use intra-group similarities minimum, the similarity between groups maximum to identify entity names. In those methods sometimes, lexical resources are being used [17-18].

2. RESEARCH METHOD

NER using Groningen Meaning Bank (GMB) corpus for entity classification with enhanced and popular features by NLP applied to the dataset [19].

100 sentences randomly selected from the dataset that includes 48000 sentences. There are 318 different entity names and 2264 words in this 100 sentences data set. 141 of 318 is location name, 49 of 318 is person names,96 of 318 is organization and 32 of 318 is date names. Labels in the dataset is given in Table 1 and an example sentence is given in Table 2.

Table 1. Labels and Explain

Tag	Description
geo	Geographical Entity
org	Organization
per	Person
gpe	Geopolitical Entity
tim	Time indicator
art	Artifact
eve	Event
nat	Natural Phenomenon
O	Other

Table 2. Example Sentence

Word	POS	Tag
They	PRP	O
marched	VBD	O
from	IN	O
the	DT	O
Houses	NNS	O
of	IN	O
Parliament	NN	O
to	TO	O
a	DT	O
rally	NN	O
in	IN	O
Hyde	NNP	B-geo
Park	NNP	I-geo
.	.	O

Because of our purpose to identify, date, person and organization name, we changed geo and GPE labels to location. We changed the time label to date. In KNIME environment, Reading from CSV File process is demonstrated in Fig. 2 we made CSV reading from a file, after that changing texts to strings and finally selecting a column for processing data.

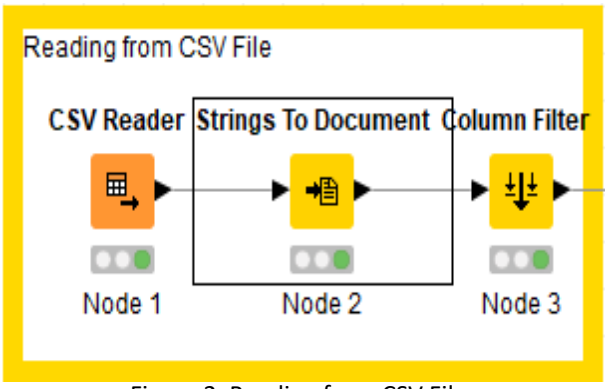

Figure 2. Reading from CSV File.

After the choosing column, some pre-processing must be done, this process is demonstrated in Fig. 3, these processes are for cleaning the punctuation, words which are shorter than a specific number of characters and stop words which are given in Table 3.

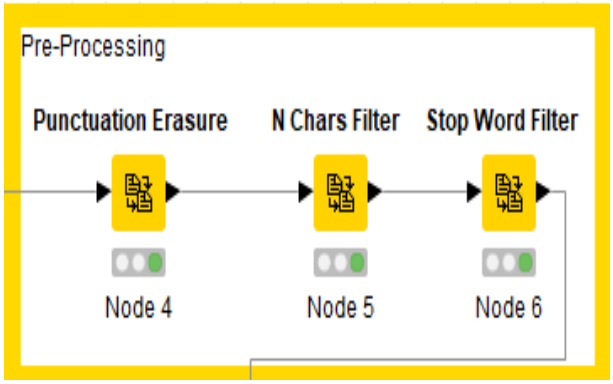

Figure 3. Pre-Processing Stage.

Table 3. English Stop Words Examples

Words
about, across, before, could, down, each, find, however, it, me, next, other, see, take, under, was, with, yet etc.

In this work, NER is performed and one method is suggested and results are discussed for assignment to unlabeled name entities by using OpenNLP library with the help of KNIME program in the data set.

KNIME name is a short version of, Konstanz Information Miner (KNIME) is an open source cross-platform data analysis, reporting integration platform. KNIME includes components of machine learning, data mining based on modular data line concept, and these tools name are node and visualization, modeling and data analysis.

OpenNLP is a library which supports many languages developed by Apache for natural language processing.

Data labeled with the help of OpenNLP which area preprocessed on KNIME environment. In this work, we focused on 4 different entity names (person, location, organization, date). Tagging stage is demonstrated in Fig. 4.

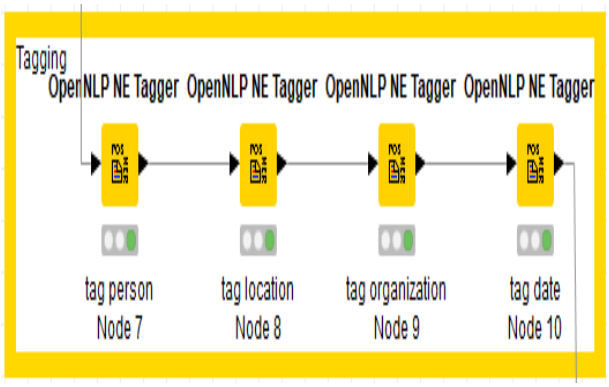

Figure 4. Tagging Stage.

After labeling 4 different entity names with the help of mode which made in OpenNLP library. As seen in Fig. 5, labeled entity names are filtered and turned to text.

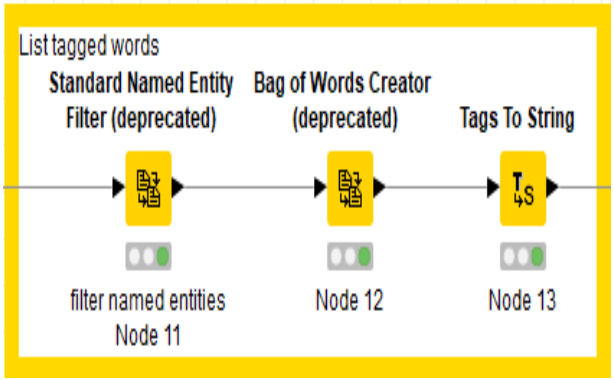

Figure 5. List Tagged Words Stage.

Filtered and turned to textualized entity names are grouped by term frequencies and term number turned to singular. This process is demonstrated in Fig. 6.

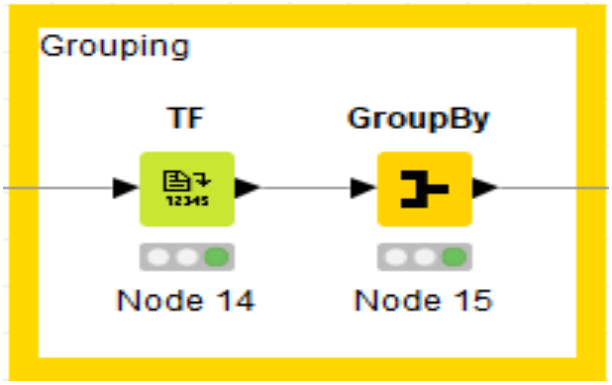

Figure 6. Grouping Stage.

After processing, text turned to a text which is labeled with 4 entity names. For visualization, the results of these processes, our grouped entity names labeled with different colors and they are showed as a cloud structure

Selected colors are demonstrated in Fig. 7, the related stage is in Fig. 8 and Cloud structure is in Fig. 9.

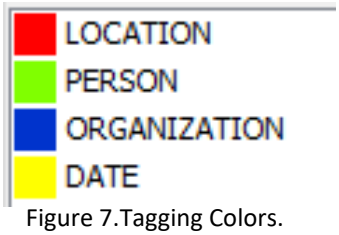

Figure 7.Tagging Colors.

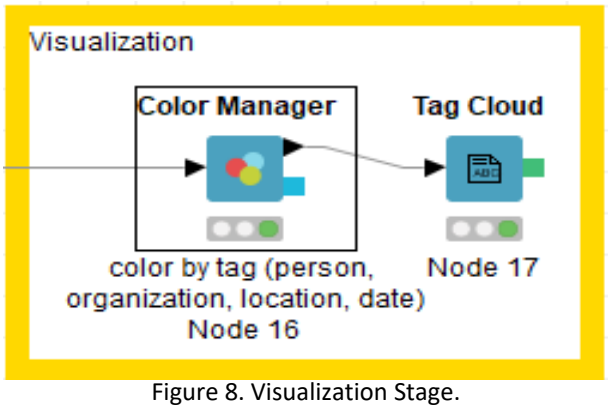

Figure 8. Visualization Stage.

Figure 9. Tag Cloud.

3. RESULTS AND DISCUSSION

In the conducted work a data set is used that includes 100 sentences 2264 words and 318 different entity names. Person, location, organization and date names labeled and they are calculated by accuracy, precision, recall, F-Score, and Kappa. 101 name labeled from 318 different names. The confusion matrix is given in Table 4 and the results calculated by metrics are given in Table 5. Metrics which we use to define trial outcomes and which are also preferred in similar studies [20-22] are Accuracy, Precision, Recall, F-Measure and Kappa statistics and are presented in Equation (1-5).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F - Measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

Kappa statistic is also used to define the classification performance which holds the following terms: P(A) reflects the agreement percentage in between classifier and underlying truth P(E) is the chance of agreement.

$$Kappa = ((P(A) - P(E))|(1 - P(E)) \quad (5)$$

Table 4. Confusion Matrix-1

	Location	Person	Organization	Date
Location	32	0	4	0
Person	0	12	0	0
Organization	1	2	18	0
Date	0	0	0	32

Table 5. Results-1

Metric	Result
Accuracy	93.069 (%)
Precision	0.9365
Recall	0.9112
F-Score	0.9236
Kappa	0.904

Accuracy calculated with 318 entity names are given in Table 6.

Table 6. Results-2

Named Entity	Accuracy (%)
Location	22.69
Organization	18.75
Person	24.48
Date	100
Overall	41.48

When results evaluated, we saw that the OpenNLP library labeled mostly location names. After this result automatically location name labeled to 217 unlabelled names. Related results are given in Table 7 and Table 8.

Table 7. Confusion Matrix-2

	Location	Person	Organization	Date
Location	137	0	4	0
Person	37	12	0	0
Organization	76	2	18	0
Date	0	0	0	32

Table 8. Results-3

Metric	Result
Accuracy	62.579 (%)
Precision	0.601
Recall	0.805
F-Score	0.688
Kappa	0.39

4. CONCLUSION

In this work, NER processed on English sentences. The used data set includes 100 sentences 2264 words and 318 different entity names. Labeling processed for 101 entity name in this work which is made on KNIME environment with the help of the OpenNLP library. Labeling rate of entity name is about 31.4 (%) even though the accuracy of labeled entity names rate's average value is 93.069 (%). Besides the rate of accuracy, precision, recall, F-score and Kappa metrics results calculated as high values. Person and Date entity names labeled correctly, but there is 11 (%) rate of error for location and 14(%) rate of error for the organization.

Next studies our purpose is to work on better methods for labeling 217 entity names which are not labeled by OpenNLP, and we will study Turkish texts.

REFERENCES

[1] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30.1 (2007): 3-26.

[2] https://imanage.com/blog/named-entity-recognition-ravn-part-1 - Online : 15.02.2018.

[3] Eken, Beyza. Named entity recognition on Turkish short texts. Master Diss.Graduate School of Natural And Applied Sciences, 2015.

[4] Rau, L.F., "Extracting Company Names from Text", In Proc. Conference on Artificial Intelligence Applications of IEEE, 1991.

[5] Chiu, Jason PC, and Eric Nichols. "Named entity recognition with bidirectional LSTM-CNNs." Transactions of the Association for Computational Linguistics 4 (2016): 357-370.

[6] Derczynski, Leon, et al. "Analysis of named entity recognition and linking for tweets." Information Processing & Management51.2 (2015): 32-49.

[7] Santos, Cicero Nogueira dos, and Victor Guimaraes. "Boosting named entity recognition with neural character embeddings." arXiv preprint arXiv:1505.05008 (2015).

[8] Marrero, Mónica, et al. "Named entity recognition: fallacies, challenges and opportunities." Computer Standards & Interfaces35.5 (2013): 482-489.

[9] Shaalan, Khaled. "A survey of arabic named entity recognition and classification." Computational Linguistics 40.2 (2014): 469-510.

[10] Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R., “A High-Performance Learning Name-finder”, In Proc. Conference on Applied Natural Language Processing, 1997.

[11] Sekine, S., “Description of the Japanese NE System Used For Met-2”, In Proc.Message Understanding Conference, 1998.

[12] Borthwick, A., Sterling, J., Agichtein, E., Grishman, R., “Description of the MENE Named Entity System as used in MUC-7”, In Proc. Seventh Message Understanding Conference, 1998.

[13] Asahara, M., Matsumoto, Y., “Japanese Named Entity Extraction with Redundant Morphological Analysis”, In Proc. Human Language Technology Conference – North American chapter of the Association for Computational Linguistics, 2003.

[14] McCallum, A., Li, W., “Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons”, In Proc. Conference on Computational Natural Language Learning, 2003.

[15] Minkov, E., Wang, R., Cohen, W., ”Extracting Personal Names form e-mail: Applying Name Entity Recognition to Informal Text”, In Proc. Human Language Technology and Conference on Empirical Method in Natural Language Processing, 2005.

[16] Brin, S., “Extracting Patterns and Relations from the World Wide Web”, In Proc.Conference of Extending Database Technology. Workshop on the Web and Databases, 1998.

[17] Alfonseca, E., Manandhar, S., “An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery”, In Proc. International Conference on General WordNet, 2002.

[18] Cucerzan, S., Yarowsky, D., “Language independent named entity recognition combining morphological and contextual evidence”, In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999.

[19] https://www.kaggle.com/steelrat11/ner-project/data - Online 15.01.2018

[20] Khanna, S. and Agarwal, S., “An Integrated Approach towards the prediction of Likelihood of Diabetes,” Proc. Machine Intelligence and Research Advancement, Katra, India, pp. 294-298, 2013.

[21] Mahardhika, Y. M., Sudarsono, A. and Barakbah, A. R., “An implementation of Botnet dataset to predict accuracy based on network flow model,” Proc. Knowledge Creation and Intelligent Computing, Surabaya, Indonesia, pp. 33-39, 2017.

[22] Raut, M. Y. and Barve, S. S., “A semi-automated review classification system based on supervised machine learning,” Proc. 1st. Intelligent Systems and Information Management, Aurangabad, India, pp. 127-133, 2017.