ISSN 2278-3091
Volume 9, No.3, May - June 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse190932020.pdf
https://doi.org/10.30534/ijatcse/2020/190932020

An Ensemble Technique for Named Entity Recognition using Conditional Random Fields and L-BFGS Optimization Algorithm

CH.Susmitha[1], D.Haritha[2]

[1]Student, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Guntur District, Andhra Pradesh, India, susmithachinnam@gmail.com

[2]Professor, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Guntur District, Andhra Pradesh, India

ABSTRACT

This paper is regarding Named Entity Recognition (NER) for English. It is a computational phonetic under-taking wherein we generally tend to glance to rearrange each word in an exceedingly record, as tending to be categorized principally into 4 classes Person, area, Organization, and name-other (Date, time and so forth). Now, we tend to begin our assessments in building a CRF (Conditional Random Fields) that found along Noun Tagger Trained with a physically labeled information of around 48,000 sentences. This noun tagger has given an associate F-score of 85%. A CRF based NER framework is then developed for English and tried it on some informational indexes. We have got F-scores among 80% to 92% in numerous preliminaries. Named entity recognition could be a crucial task for a number of NLP applications.

Key words: NER, Conditional Random Fields, L-BFGS.

1. INTRODUCTION

Named entity recognition (NER) additionally stated as entity lumping/extraction, a notable strategy utilized in information extraction to acknowledge and segment the named entities and arranges (or) classifies them below predefined classes.NER became initially used in Message understanding conference-6(MUC-6) in1995.

Many ambiguities emerge due to the nature of linguistic communication. There are particular terms in any text document that constitute precise entities which can be greater informative and have an identical context, which mainly refers to the text that represents real-world objects like places, people, organizations, and then on that are frequently meant by way of appropriate names, these entities are likewise as named entities.

NER may be carried out in many ways. NER models are created using rule-based technique via making use of statistical models, (i.e., machine learning approach) or hybrid methodology that combine both rule based and machine learning. Rule based technique depends on linguistic rule. So, it is very hard to evolve into different languages and resources unique to a language utilized in Rule-based approaches are cannot be convenient to other languages.

In ML procedures, a statistical approach work using annotated corpus as training data and builds a probabilistic model with the elements of the information like rule-based techniques. The corpus labeled with named entities is discovered to produce the features of the data, and then it is accustomed to calculating and picks the most probable NE's.

In this paper, we will represent our new NER framework for the English language based on CRF using annotated corpus data for named entity recognition. We would be utilizing sklearn-crfsuite to develop our NER together with eli5and L-BFGS algorithm for optimization and obtaining model parameters. Moreover, this work also supports information extraction, text summarization, etc.

The paper is arranged according to the accompanying: In Sect 2 the related work is provided. Conditional random fields and L-BFGS are explained within the following Sect 3. Sect 4 gives a brief explanation about the corpus, English language, and characteristics of NE in language.

The Experiment and evaluation are demonstrated in Sect 5. Results are represented in Sect 6 and conclusion in Sect 7.

2. RELATED WORKS

Given that in the21st century, the web information is growing steadily, and an enormous amount of knowledge is increasing exponentially. For extracting that data and processing that information many tools and technologies became essential. Underneath, such circumstances many numerous advancements appeared for information retrieval, extracting information and computational linguistics, etc., NER is one in each of the fore most Vital branches of natural language processing.

NER performs a crucial role in the extraction and retrieval of data, machine translation, and text summarization soon. Many research works are performed on NER. The approaches of NER are divided into a Rule-based approach and machine learning-based approaches. Rule based approach is completed manually and this form of approach is tedious and costly. The Machine learning-based approach received more attention as large NE tagged corpora is accessible.

Hsu Myat Mo (&), Khin Thandar Nwet, and Khin Mar Soe performed ner using conditional random fields (CRF) and mentioned some experimental results on Myanmar language [1].Gowri Prasad presents a way to modify named entity recognition for English and Hindi Languages using maxi-mum entropy-based models, CRF and SVM [2].

Michal Konkal and Miloslav konopik designed the NER system segment representation of multi-word entities for four languages particularly English, Spanish, Dutch, and Czech using maximum entropy and CRF model [3]. Vijay Krishna R and Sobha L represented a domain focused Tamil Named Entity Recognizer of the touristry domain. Named entities with a class-conscious tag set containing 106 tags are handled throughout this paper; they build Conditional random fields (CRF) via training the noun phrases of the training data [4].

3. CONDITIONAL RANDOM FIELDS

The CRF's model is anticipated by Lafferty in 2001[6] is a typically discriminated likelihood non-directional Chart model supported by entropy model and HMM that focuses on serial labeling. CRF's area unit rudderless graphical models a selected instance of that compares to restrictively prepare limited/finite state machine based mostly same exponential kind as most entropy model. CRF's have demonstrated exact accomplishment in POS tagging, phrase tagging, and CWS [5].

Conditional models are accustomed to labeling the observation sequence x by choosing the label sequence y that maximizes the conditional probability p(y|x).The conditional idea of those models implies that no exertion is squandered on demonstrating the perception and one is liberated from making undesirable assumptions about these arrangements. Conditional random fields maintain a strategic distance from the name inclination issue and coordinated graphical models are upheld by most extreme entropy models [7].

Lafferty [7] proposed probability of a selected label sequence y has given observation sequence x to be normalized product of potential functions, each of the shapes as:

$$p(y|x,\lambda) = \frac{1}{\mathrm{Z(x)}}\exp\left(\sum_{\mathrm{j}} \lambda\,\mathrm{jFj(y,x)}\right)$$

3.1. L-BFGS Algorithm

In this paper, we are using the L-BFGS algorithm for gradient descent for optimization and model parameters in the own family of Quasi-Newton. L-BFGS is an optimization algorithm that approximates the Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS).

The algorithm minimizes f(x) over unconstrained estimations of the real vector x where f is a differentiable scalar function. To influence its search through variable space L-BFGS uses an estimate of Inverse Hessian matrix, where as BFGS stores n*n approximation to the inverse Hessian.

We are additionally making use of L1 and L2 (Coefficient of Lasso and Ridge Regularization).

We have to attenuate f(x) using a second-degree approximation method. Taylor arrangement of the function f(x) is given below:

$$f(x_0 + \Delta x) = f(x_0 + \nabla f(x_0)^T \Delta x + \frac{1}{2}\Delta x^T . H\Delta x$$

Where delta f is gradient function and H is hessian.

Taylor series of gradient function:

$$\nabla f(x_0 + \Delta x) = \nabla f(x_0) + H\Delta x$$

To solve the equation, we must find the minimum:

$$\nabla f(x_0 + \Delta x) = 0$$

From here:

$$\Delta x_0 = -B^{-}1 . \nabla f(x_0)$$

Now, Hessian should be appointed there is a special way of approach for every method belonging to the current family and appointing it.

3.2. The Contrast between BFGS and L-BFGS

L-BFGS requires lesser memory than the standard BGFS and works with a large dataset. Hessian inverse matrix is employed in both the algorithms to control variable space searching.

BFGS stores a dense n*n approximation to the inverse Hessian, where as just hardly a vectors are stored in L-BFGS which represent the guess implicitly. So the memory is saved by using this approach.

4. CHARACTERSTICS OF ENGLISH LANGUAGE

ENGLISH LANGUAGE is the authentic language in 67 completely different countries everywhere on the planet. In the English language, we have eight components of Speech named as Noun, Pronoun, Verb, Adjective, Adverb, Preposition, Conjunction, and Interjection.

The Parts of speech suggests how the word used within the sentences and its importance even as linguistically. Receptiveness, Heterogeneousness, Simplicity of Inflexion, Fixed Word Order, Growth of Intonation, Use of Periphrasis are the most traits of the English language.

The English language has many ambiguities in both syntactic and semantic meanings. Many kinds of research are on going to solve these ambiguities however the effective approach isn't introduced nevertheless as a result of the language evolves the ambiguities also besides arise.

4.1. Data Preparation &Corpus

Conditional Random field is trained on a sequence of input data to analyze the transformation from one label to another. To empower such an algorithm, we need to define some features for different transitions. To transform every word into a feature, we use **word2feature ( )** function depicting following the features or attributes:

- Lower case of the word.
- Suffix containing the last 3 characters.
- Suffix containing the last 2 characters.
- Flags to determine capitalized, title-case, numeric data, and POS tag.

To decide the start of the sentence or end of a sentence, we attach attributes associated with previous and next words of the sentence.

The data presented in this paper is taken from Groningen Meaning Bank. The GMB dataset utilizes IOB tagging or within, Outside Beginning. IOB is a common format for tagging tokens.

- **I-prefix** Inside the chunk for the word.
- **B-prefix** Beginning chunk for the word.
- **O-Tag** Outside of the chunk.

Anything outside the classes is denoted by **O**. The NER tag set used in the corpus is illustrated below:

PER	Person
GEO	Location
ORG	Organization
GPE	Geo−Political Entity
EVE	Event
TIM	Time-indicator
ART	Arte fact
NAT	Natural Phenomenon

The annotated corpus data for named entity recognition has around 48000 sentences where every sentence is tokenized and assigned with POS tag and NER tagging.

We have a total of 35178 words 42 POS tags and17 NER tags in distinctive tagging for the entire corpus is given in the table 1.

Table 1: NER Tagging

B—geo	37644
B—tim	20333
B—org	20143
I—per	17251
B—per	16990
I—org	16784
B—gpe	15780
I—geo	7414
I—tim	6528
B—art	402
I—art	297
I—eve	253
B—nat	201
I—gpe	198
I—nat	51

4.2. NER Process

The first step is pre processing. Pre-processing of the original text includes word segmentation continued with POS tagging so that alternative characteristics of the original text are expressed explicitly. To accomplish the content of word division and POS tagging we have used the LTP platform.

The tagging of the training set is done in a very manual manner to Mark the entities. A word is chosen from the text corpus segmentation, use IOB annotation method and find the preparation set element explanation arrangement where I (Internal) signifies the internal tagging of the chunk, B (begin) denotes the starting of the entity and O(other) indicates words, punctuation, etc., Named entity and its annotation are shown in the table 2.

Table 2: Named Entity and Annotation

Word	POS	NER Tag	Word	POS	NER Tag
On	IN	O	plans	VBZ	O
Tuesday	NNP	B—tim	To	TO	O
Iranian	JJ	B—gpe	install	VB	O
President	NNP	B—per	6000	CD	O
Mahmoud	NNP	I—per	new	JJ	O
Ahmadine jad	NNP	I—per	centrifuges	NNS	O
Announced	VBD	O	At	IN	O
Tehran	NNP	B—gpe	Nantaz	NNP	B—geo

5. EXPERIMENT AND EVALUATION

The Experiment and Evaluation are completed on the GMB dataset using the CRF toolkit. There are a total of 47,959 sentences including 35178 distinctive words within the corpus. This corpus has been used as training data for the CRF-based NER system.

In NER, if the recall proportion is high then it indicates that the named entity recognition is additional, thus along these lines making naming substances increment, it is conceivable to accuracy.

Confusion Matrix is used to derive an f1 score. It maintains stability between Precision and Recall.

$$F1\,score = 2\frac{Precision \times Recall}{Precision + Recall}$$

Now we will evaluate our model execution for NER tagging on the test data. We use preferred classification metrics like Precision, Recall, and F-1 score to measure the model accuracy. We obtained a precision of 86%, recall of 85% and f1-score got 85%. We can see the whole CRF metrics for NER tagging in the below table 3.

Table 3: CRF metrics for NER tagging

NER	Precision	Recall	f 1—score
B—org	0.81	0.73	0.77
B—per	0.85	0.84	0.84
I—per	0.85	0.90	0.88
B—geo	0.86	0.91	0.89
I—geo	0.81	0.80	0.81
B—tim	0.93	0.89	0.91
I—org	0.82	0.79	0.80
B—gpe	0.97	0.94	0.96
I—tim	0.84	0.81	0.82
B—nat	0.50	0.24	0.32
B—eve	0.51	0.33	0.40
B—art	0.36	0.14	0.20
I—art	0.24	0.07	0.10
I—eve	0.45	0.19	0.27
I—gpe	0.86	0.53	0.66
I—nat	0.57	0.22	0.32
Microavg	0.86	0.85	0.86
Macroavg	0.70	0.58	0.62
Weightedavg	0.86	0.85	0.85

O tag is left out to evaluate the model performance on the remaining tags. We have obtained an F-1 score of 85% and also the statistics show that the model has learned the transitions quite well. By feature engineering and hyper-parameter tuning, we can achieve better results.

6. RESULTS

To calculate the correctness of the model, we prepared a sample document and evaluated the model performance on it. We have got to build a data pipeline for the model.

Firstly, we tokenized the text followed by POS tagging then we have extracted the features from the POS tagged text document and used sent2features. Later we used the CRF model to predict the features then combined the text tokens with NER tags and retrieved applicable named entities from it. The results are displayed in the below table 4.

Table 4: NER Results

Entity	Tag
March22	I—tim
Indian Journal of Medical Research	I—org
ICMR	B—org
Balaram Bhargava	I—per
April-10	I—tim
Kismayo	B—geo
Somalia	B—geo
Lower Juba	I—geo

7. CONCLUSION

Named entity recognition plays a very significant role in NLP for automated information extraction. The main aim of NER is extracting the information from the text as in today's world a huge amount of information is available. This paper is intended to approach Named Entity recognition in a statistical way using the CRF model and L-BFGS algorithm. For this paper, we used a ML approach for getting higher accuracy and fast processing as the data is huge. A lot more analysis would be done in future work to include more accuracy.

REFERENCES

[1] Hsu Myat Mo, Khin Thandar Nwet, and Khin Mar Soe, **CRF-Based Named Entity Recognition for Myanmar Language**, Springer International Publishing AG 2017 J. Pan et al. (eds.), Genetic and Evolutionary Computing, Advances in Intelligent Systems and Computing 536, DOI10.1007/978-3-319-48490-7_24

[2] Gowri Prasad and Fousiya K.K, **Named Entity Recognition approaches**, A Study applied to English and Hindi Language. 2015, IEEE. https://doi.org/10.1109/ICCPCT.2015.7159443

[3] Michal Konkol and Miloslav Konopik, **Segment Representations in Named Entity Recognition**, Springer International Publishing Switzerland 2015 P.Kr´al and V.Matousek (Eds.):TSD 2015, LNA I9302, pp.61–70,2015.DOI:10.1007/978-3-319-24033-67.

[4] Vijayakrishna, R., Sobha, L., **Domain focused named entity recognizer for Tamil using conditional random fields**, In: Proceeding of the IJCNLP-08 Workshop on NER for South East Asian Languages, January (008).

[5] McCallum, A., Li, W ., **Early results for named entity recognition with conditional random fields, feature induction and web enhanced lexicons**, In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL-Volume 4. Association for Computational Linguistics, vol. 2003, pp. 188–191 (2003) https://doi.org/10.3115/1119176.1119206

[6] Zhao,H. ,Huang,C.N., Li,M, **An improved Chinese word segmentation system with conditional random field**, In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: July 1082117 (2006).

[7] Lafferty, J., McCallum, A., Pereira, F, **Conditional random fields: probabilistic models for segmenting and labeling sequence data**, In ICML 2001 Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289 (2001).

[8] Darvinder Kaur and Vishal Gupta, **A survey of Named Entity Recognition in English and Indian Languages**, IJCSI International Journal of Computer Science Issues, Vol.7, Issue6, November 2010.

[9] Sanjana kamath and Rupali Wagh, **Named Entity Recognition Approaches and Challenges**, International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 2, February 2017

[10] Abhijit Debbarma, Paritosh Bhattacharya, Bipul Shyam Purkayastha, **Named Entity Recognition for a Low Resource Language,** International Journal of Recent Technology and Engineering(IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019 https://doi.org/10.35940/ijrte.B2085.098319

[11] Shengli Song, Nan Zhang and Haitao Huang, **Named entity recognition based on conditional random fields**, Springer Science + Business Media, LLC 2017 https://doi.org/10.1007/s10586-017-1146-3