IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 52-56

A Study of Named Entity Recognition Systems in Languages

Puneet Sharma[1], Bramah Hazela[2], Deepak Arora[3], Richa Chaturvedi[4]
Department of Computer Science & Engineering,
Amity School of Engineering and Technology,
Amity University, Uttar Pradesh Lucknow Campus, INDIA

Abstract: *The overall meaning of any elaborated text can be known by identifying some entities present in text; hence Named Entity Recognition plays a very crucial role. Detection of named entities is one of the most crucial task so in the process of message understanding. This paper focuses on the encoding schemes used by NER and the languages that NER systems have been applied to. This paper also introduces a methodology that can be used to perform sequence labelling using Conditional Random Fields.*
Keywords*: BIO · HMM · MEMM ·CRF · NER ·LBFGS · Recall*

I. Introduction

The NER issue has gotten much consideration, as NER frames the essential building piece of any Information Extraction framework. Despite the fact that classifying the correct named entity exceptionally difficult errand in English, the task benefits a great deal from the distinguishing orthographic feature of capitalization. At the point when this component is absent, as in capitalized message, or is available toward the beginning of a sentence, equivocalness increments, and requires more information sources to determine the issue.

The possibility of automatic extraction of imperative data from content records originates from the season of first ventures in the normal dialect handling. Its significance quickly develops with the ascent of the advanced news, online networking, and social media and so on. The measure of data is overpowering and data extraction can oversee it. Named Entity Recognition is a basic subtask of data extraction. It tries in perceiving and characterizing multiword articulations with uncommon significance, e.g. people, associations, areas, dates, and so on.

II. Literature Survey

BIO model can also be called as IOB model comes into play when we want to discard any problems related to sequencing. The BIO stands for beginning, inside and outside. BIO model is considered to be the industry standard. It divides the tag into two parts B_X and I_X which means beginning of tag and continuation of tag, respectively. There is also a type "O" which means not a tag. In the above mentioned example the sentence will be labeled as "This /O year /O, /O I /O am /O planning /O to /O visit /O Costa /B-LOC Rica /B-LOC during /O my /O vacations /O." It uses two different forms of representing the same thing. Bio-2 will use the Beginning tag for the very first word that is present in the entity, while the BIO-1 uses the first words tag if the condition that the following entity is of the similar type. As stated by Michal Konkol and Miloslav Konopiık, The conferences held by CoNLL in 2002 and 2003 used the BIO encoding for all the annotations and since then this type of encoding has been used over all the other in the sector of Named Entity Recognition.

Michal Konkol and Miloslav Konopiık concluded in their study that BIO does indded performs much better than any of its alternate forms like BILOU. Infact the performance of BILOU was very poor when it was tested with the English corpus using the CRF model. A somewhat similar result was found by Lev Ratinov and Dan Roth.

2.1 Ner Systems Applied On Languages

NER has been applied to many languages like Hindi, Gujrati, Bengali, Chinese and Russian. Chirag Patel and Karthik Gali in their paper made a Part of Speech of Tagging for Gujrati language. They used a CRF model. The features given to CRF are appropriately picked keeping the semantic part of Gujarati as a top priority. As Gujarati is presently a less favoured dialect in the feeling of being asset poor, physically labelled information is just around 600 sentences. The tag set contains 26 extraordinary labels which is the standard Indian Language (IL) tag set. Both labelled (600 sentences) furthermore, untagged (5000 sentences) are utilized for learning. The model had accomplished an accuracy of 92% for Gujarati writings where the training dataset is of the thousand words and the test dataset is of five thousand words. It utilizes a CRF to factually label the test corpus. The CRF is prepared utilizing features over labeled and untagged information. A CRF at the point when furnished with great features gives exactness much superior to anything different models. The instinct here is

that on the off chance that we change over the semantic standards particular to Gujarati in to features gave to CRF; at that point we make utilization of points of interest of both factual and run the show based approach. However, because of absence of control and adaptability not all features be fused in the CRF. So after the CRF is done we do the blunder investigation. From the blunders we detail rules, which are general and dialect particular, and after that change over them to new features and apply them back to CRF. This expands the exactness.

In another research conducted by Mozharova V. A. and Loukachevitch N designed an NER system for the Russian language using the CRF model. Current machine-learning ways to deal with data extraction regularly incorporate features in light of extensive volumes of information in shape of clusters of words. They tested their model on the open Russian dataset called "People 1000" tagged with individual names. The authors also tagged the dataset with entities like location, medias, organization. They additionally showed the comparison between the two types of encoding: IO encoding and IOB encoding. They achieved an F-score of 95.63% which was slight less compared to rule based system which had the F score value of 96.62%.

Yuejie Zhang et al, in their paper exhibit the capacity of Conditional Random Field (CRF) consolidating with different features to perform powerful and precise Chinese Named Entity Recognition. They have depicted the different component formats including neighborhood feature templates and global feature template used to extricate numerous features with the assistance of human information. In addition, they demonstrated that human information can effectively smooth the model and thus the need for training data for CRF may be decreased. The research conducted on People's Daily dataset showed that their model is a viable example to join statistical model and human knowledge. Another experiment that was conducted on a different corpus affirms the above conclusion, which demonstrates that features were consistence on various testing information.

Study conducted by Asif Ekbal and Sivaji Bandyopadhyay reports about the improvement of a Named Entity Recognition (NER) framework in Bengali by joining the outputs that was obtained from the two classifiers, to be specific Conditional Random Fields (CRF) and Support Vector Machines (SVM). Lexical context pattern, which are created from an unlabelled corpus of 10 million wordforms in an unsupervised way, have been utilized as the features of the classifiers with a specific end goal to enhance their execution. The authors have post-processed the models by considering the second best tag of CRF and class splitting procedure of SVM with a specific end goal to enhance the execution. At last, the classifiers are consolidated together into a last framework utilizing three weighted voting procedures. The results obtained demonstrate the viability of the proposed approach with the general normal review, exactness, and f-score estimations of 91.33%, 88.19%, and 89.73%, respectively.

2.2 CRF vs. HMM and MEMM

Sequence labelling alludes to the supervised learning task of giving out a label to every component of a sentence. An example of this is Part-of-Speech labelling, NER and Gene Prediction as mentioned by Allen et al. 2004 and Allen and Salzberg 2005. In such cases, a particular tag can't be considered to be isolates from the overall meaning or context (i.e. the former and succeeding components of the sentence and tags corresponding to them). Two of the most prominent arrangement models are Hidden Markov models (HMM) which were proposed by Rabiner and Conditional Random Fields (CRF) (Lafferty et al. 2001). Because of the generally high dimensional feature spaces (specially in the case of CRFs), these models regularly require a lot of labelled data to be effectively trained, which ruins the development and making of datasets and makes it relatively restrictive to do with a solitary annotator. In spite of the fact that in a few areas, the utilization of unlabelled data can help in making this issue less serious, a more regular arrangement is to depend on various annotators. For instance, for some assignments, AMT can be utilized to mark a lot of information. Be that as it may, the substantial numbers expected to adjust for the heterogeneity of annotators ability quickly raise its genuine cost beyond acceptable quality. A closefisted arrangement should be composed that can manage such genuine limitations and heterogeneities.

Regardless of the several of methodologies exhibited for learning models, the issue of sequence labeling from numerous annotators was left basically untouched, with the main significant work being the work by Dredze et al. The creators propose a strategy for learning organized indicators, in particular Conditional Random Fields. This is accomplished by changing the CRF target function utilized for training purpose through the incorporation of pre-label prior. The per-tag priors are then re-evaluated by making use of their probabilities under the entire corporas. Along these lines, the model is fit for utilizing learning from different parts of the dataset to lean toward specific marks over others. By emphasizing between the calculation of the normal estimations of the name priors and the estimation of the model parameters, the model is required to offer inclination to the less uproarious names. Consequently, we can see this procedure as self-preparing, a procedure whereby the model is prepared iteratively alone yield. In spite of the fact that this approach makes the model computationally tractable, their trial comes about demonstrate that the strategy just enhances execution in

situations where there is a little measure of preparing information (low amount) and when the names are uproarious (low quality).

But, classifying every label independently might be easier, because there is a certain disadvantage faced by raw labeling where single label has to be assigned to every element of the sentence. Though classifying every label independently seems to be simple, but has seen to be performing quite effectively in many scenarios. While identifying the Part of speech tags it is almost very difficult to have a determiner which is just preceded by a verb, there having knowledge about the locally sequencing words can be more efficient. One way this can be done is by applying the HMM model. The HMM serves this problem by having two different probabilities one for transitional distributor and another for emitted distributer which means the transitional distributor will determine the likeliness of presence of a proverb followed by a determiner and emitted distributer will determine the likeliness of presence of the word " before" given that this very word is a determiner. If these transitions are locally present, Viterbi algorithm will execute perfectly during the test time. This is known as the Markov Assumption. But the reason of not using the Hidden Markov Model is that the emitter probability is of the type of P (word | entity), while a model similar to P (word | entity) would be more welcomed. The only reason because of this preference is that it can support a plethora of overlapping feature. As mentioned by Lev Ratinov and Dan Roth the global inference over the second-order Hidden Markov Model feature does not captures the non-local properties of the task.

In the experiment conducted by Radu Florian et al, the Hidden Markov Model classifier is used. Sequence labelling is conducted by giving every word either one entity type of Not a Name which denotes marks that word to not belonging to any named entity. The various transitions in the Hidden Markov Model is divided into distinct spaces, where every space represents every named entity type and another space represents a Not a named entity tag. Within every space the likelihood of the words appearing to be belonging to that region is present.

Radu Florian conducted an experiment where the combined various classifier for NER system. The classifiers combined together were robust linear classifier, maximum entropy, transformation-based learning, and hidden Markov model. They were grouped in various conditions. It was tested on the English corporus and when no other training resource was utilized, the performance value of this combined classifier was up to 91.6 %.In another research paper Radu Florian in 2002 presented a method of improving the classifiers performance by staking on different approaches. He stacked two algorithms, Transformation-based learning, sparse network of winnows or also known as Snow (Muñoz et al., 1999) and a forward-backward algorithm. The result from one classifier becomes the input of the succeeding classifier.

III. Methodology

We made use of the CRF library is CRFSuite which was developed by Naoaki Okazaki using C/C++. The library is as of now simple to utilize given its order line interface. Pycrfsuite is accessible for utilizing the API in Python. This Python module is precisely the module utilized as a part of the POS tagger in the nltk module. CRFsuite is an execution of Conditional Random Fields (CRFs) which is used for tagging data in sequential form. It provides many features like-

a) Quick training and labeling - The essential mission of this product is to prepare and utilize CRF models as quick as would be prudent.
b) Straightforward format- The information design is like those utilized as a part of other machine learning instruments; each line comprises of a mark and characteristics of a thing, consecutive lines speak to a grouping of things (a void line signifies a finish of thing succession). This implies clients can plan a subjective number of features for everything, which is inconceivable in CRF++.

- CRFsuite executes:
- Limited memory BFGS (L-BFGS)
- Stochastic Gradient Descent (SGD)
- Averaged Perceptron
- Adaptive Regularization Of Weight Vector (AROW)

CRFsuite can yield precision, recall and F1 scores of the model assessed on test information. A productive record organizes for putting away/getting to CRF models utilizing Constant Quark Database (CQDB). It requires a little investment to fire up a tagger since a planning is done just by perusing a whole model document to a memory square. Retrieving the heaviness of an element is additionally brisk. C++/SWIG API. CRFsuite gives a simple to-utilize API for C++ dialect (crfsuite.hpp). CRFsuite likewise gives the SWIG interface to different dialects (e.g., Python) over the C++ API. See the API Documentation for more data.

For training of the NER model, we require some labelled corpora. The dataset that was be utilized was 500 news gold standards; it consists of 500 English news articles from online news platforms. This dataset is

already labelled. It is an English corpus in the NLP Interchange Format (NIF). We used the XML version of the dataset.

The first step is preparing the dataset for training. So we have to tag or classify every entity of the sentence into two possible classes, either irrelevant or part of a named entity. We used the BeautifulSoup library for this purpose. After these steps parsing of the data has been done.

Now comes the phase of training of the CRF model, for this we will have to create feature for the entities in the sentence. Part of Speech tagging's majorly used as the feature.

Creating Features

Given the POS labels, we would now be able to keep on generating features for every token in the corpora. The features that will be helpful in the training process rely upon the case we are currently working on. The following are a few of the features for entity e in named entity recognition:

a. The entity e itself (changed over to lowercase for standardization)
b. The prefix/postfix of e (e.g. - ing)
c. The words neighbouring the entity
d. whether w is in capitalized or lowercase
e. whether w is a number, or contains digits
f. The POS tag of w, and those of the neighbouring words
g. entity contains a unique character (e.g. hypen)

We used the function word2feature () for extracting the feature for our dataset.

Now for training the model, we have to first set up the training data so that it corresponds to tagged labels. Likewise, to check the exactness of the model checked, we parted the whole dataset into two parts training dataset and testing dataset. For this we used train_test_split() function in scikit-learn.

In pycrfsuite, a CRF model can be trained by first making a trainer, and afterward submitting the training data and corresponding labels. After that, set the parameters and call train () to begin training the process. With comparatively smaller datasets in our case, the training with max_iterations=200 can be done in almost no time.

Now after having the trained model and we applied it on our test data to see whether it gives sensible outcomes or not. Then we saved the model in a document named crf.model.

To consider the effectiveness of the CRF tagger prepared above in a more quantitative manner, we can evaluate the precision, recall and F1 score on the data used for testing purpose.

IV. Conclusion & Future Works

CRF model is a best in class for sequential tagging, which can utilize the features that are used everywhere in a more adequate and successful manner. In this paper, we have proposed and executed the CRF-based NER. Many researches has demonstrated that the CRF model beats the other machine learning techniques, like SVM, MEMMS etc. in the task to named entity recognition and classification. Authors have compared and studied the Conditional Random Field or CRF and compared CRF to HMM and MaxEnt and found out the reasons which supports that CRF are indeed best to be used in scenarios where natural language processing is involved. Moreover, there are several things by which we can improve the performance, including creating better features or tuning the limitations of the CRF models. We can also try to incorporate the numerical features like, number of characters in the model.

References

[1]. Mozharova V.A., Loukachevitch N.V. (2017) Combining Knowledge and CRF-Based Approach to Named Entity Recognition in Russian. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham

[2]. Dredze, M., Talukdar, P., & Crammer, K. (2009). Sequence learning from data with multiple labels. In ECML-PKDD 2009 workshop on learning from multi-label data.

[3]. Chirag Patel and Karthik Gali, (2008)Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields, In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages

[4]. Lev Ratinov and Dan Roth (2009), Design challenges and misconceptions in named entity recognition. In: Proc. of the Thirteenth Conference on Computational Natural Language Learning. CoNLL '09, Stroudsburg, PA, USA

[5]. John Lafferty, Andrew McCallum, and Fernando C.N (2001), Conditional Random Fields: Probabilistic Model

[6]. for Segmenting and Labeling Sequence Data. In Proc. of the Eighteenth International Conference on Machine Learning. ICML, San Francisco, CA, USA

[7]. Zhang Y., Xu Z., Zhang T. (2008) Fusion of Multiple Features for Chinese Named Entity Recognition Based on CRF Model. In: Li H., Liu T., Ma WY., Sakai T., Wong KF., Zhou G. (eds) Information Retrieval Technology. AIRS 2008. Lecture Notes in Computer Science, vol. 4993. Springer, Berlin, Heidelberg

[8]. Dredze, M., Talukdar, P., & Crammer, K. (2009). Sequence learning from data with multiple labels. In ECML-PKDD 2009 workshop on learning from multi-label data.

[9]. Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang, (2003), Named entity recognition through classifier combination. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL

[10]. Michal Konkol and Miloslav Konopík. 2015. Segment Representations in Named Entity Recognition. In Proceedings of the 18th International Conference on Text, Speech, and Dialogue - Volume 9302,NY, USA