Asian Journal of Information Technology 15 (21): 4309-4317, 2016
ISSN: 1682-3915

A Machine Learning Approach to Named Entity Recognition for the Travel and Tourism Domain

Jobi Vijay and Rajeswari Sridhar
Department of Computer Science and Engineering, Anna University, Chennai, India

Abstract: This study aims to develop a Named Entity Recognition (NER) that creates new tags that facilitate fast query processing, information retrieval and data preprocessing of Travel and Tourism Domain. We have used a machine learning approach that uses domain specific knowledge to train the data and label the entities with appropriate tags. Conditional Random Fields (CRF) is implemented and used to train the input domain specific data that yields good performance. Experimental and evaluation result shows that the learned model yields a travel and tourism domain specific NER recall of 82%, precision of 85%, accuracy of 82% and F-measure of 83%. Thus learned CRF model builds a domain specific NER for tourism, travel, hotel and point of Interest domain and tags the domain keywords with appropriate tags.

Key words: Named entity recognition, statistical model, machine learning, conditional random field classifier, travel and tourism data analytics

INTRODUCTION

Travel and Tourism is an evergreen industry worldwide. According to World Travel and Tourism Council (2015), tourism is one of the major industries of all the countries that contribute to the Gross Domestic Product (GDP). As reported by Fog Computing World (2014), huge volumes of zettabytes of data are generated worldwide through internet and other sources which are evolving in digital and non-digital form. Ge *et al.* (2014) pointed out that due to the huge volume of date, during the current years there is a shift to data driven marketing. These data consists of very useful knowledge and hidden insights. Enormous data are generated in travel and tourism domain through internet, newspaper, articles, reviews about hotels and point of interest. The necessity to retrieve the required information from internet that is domain specific solves problems and can be used for fast query processing, efficient understanding of context of data, understanding the customers, improving the business and for personalized recommendation.

In order to retrieve information from this data, the data has to be cleaned and pre-processed. One of the pre-processing steps in data analysis is Named Entity Recognition (NER). According to wikipedia, "NER is a subtask ofinformation extraction that explore to discover and categorize elements in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc". It is also called as entity identification or entity chunkingor entity extraction. The best standard example for NER is Stanford's Named Entity Tagger. This Stanford NER developed by Finkel *et al.* (2005) tags the given sequence of words into name of person, organization, percent, money, data, location and time.

NER is basically used for defining entities in generic domain. The generic domain NER is not suitable for closed or domain specific NER as the characteristic changes between domains. The generic domain NER can only retrieve common information such as name of the people, places and organizations. The predefined tags and entities and tags should be modified and tuned to suit the domain specific NER. There has been information retrieval in all fields like biological, chemical, agriculture and medical industry. As basic standard NER are not helpful to retrieving domain related entities and keywords it resulted in the need for domain specific NER. Lot of research work has been carried out in the medical domain and few researches in the agricultural domain.

In the travel and tourism domain, NER is the subtask of information extraction that automatically labels the Point of interest, reviews, rating of hotels, seaso which plays a vital role than name of location, organization, time etc. The preliminary research that had been done in the area of NER are explained.

There are several methods to develop a NER such as statistical analysis, semantics analysis, knowledge base, domain specific, rule based, machine learning approach such as supervised, semi-supervised and unsupervised learning and hybrid.

Corresponding Author: Jobi Vijay , Department of Computer Science and Engineering, Anna University, Chennai, India

The research study on NER was initially based on handcrafted rules and heuristic methods published in the year 1991 by Rau (1991). MUC-6 triggered the use of NER which introduced many other NERs not only in English language but in many other languages such as Chinese, French, Greek, Italian, Hindi, Punjabi and Malayalam. The named entity recognition and Classification was introduced by Coates (1992) which further classified the entities into subcategories. Thielen (1995) also introduced the machine learning technique of classification into NER for classifying the entities.

Rule based named entity recognition system enables extraction of real life task through a set of rules which are either hand written or learnt through various real life examples. A rule based NER system basically consists of set of rules and set of policies governing the rules. There are various approaches of NER used in the literature like multiple entity recognition boundary recognition and whole entity recognition. Carreras *et al.* (2002) used the whole entity recognition which is the classic approach that is used in rule based NER. This classic NER ensures that there is no dependency between various entities and the rule is modeled based on left and right content of the keyword to be named which was proposed by Carreras *et al.* (2002) and Lee *et al.* (2003).Lafferty *et al.* (2001) and Freitag and Kushmerick (2000) proved that the rules can be applied to find out the boundaries using the left and the right context and various rules are identified to independently identity the tags using pre and post words. Rules for Multiple Entities can be used for modeling the dependency that exists between entities (Soderland, 1999).

Machine learning based NER system, the identifying suitable tags are considered as a classification problem and statistical method is used to solve it. There are many machine learning supervised approaches that had been used for Named Entity recognition such as Hidden Markov Model (HMM), decision trees, Maximum Entropy models (ME) and Support Vector Machines. These supervised models learn rules based on discriminative features. HMM model is the first machinelearning model that is used in English language for solving NER. IdentiFinder designed by Freitag and Kushmerick (1999) and Bikel (1999) used HMM Model for solving NER problem. Maximum Entropy based machine learning model is used for classification of NER by directly learning theweightage for discriminative features. The MENE system proposed by Borthwick (1999) and Curran and Clark (2003). ME tagger applied the Maximum Entropy algorithm for tagging entities. SVM was used by Paul to tackle NER as a binary decision problem.

The semi supervised machine learning algorithms that are used for NER are CRF and Bootstrapping based methods. This algorithm solves the problem of unavailability of golden standard data and sparse availability of data. CRF algorithm was later used in which learning is done based on input which is sequence of words by John. Finding NER tags using Adaboost algorithm which is a boot strapping based machine learning method was proposed in the year 2002 by Kazama *et al.* (2002) which uses BIO labeling scheme. BIO is a very popular model which is used in NER where B indicates beginning word of NE, I point the inside or intermediate word in a NE and O is the word outside the NE.

Unsupervised machine learning methods ruled out the need for large annotated corpus and it build the representation from data. KNOWITALL is a NER which uses domain independent system which extracts information from the web proposed by Etzioni *et al.* (2005). Unsupervised NER across various languages was also developed by Munro and Manning (2012).

The hybrid NER is the combination of both rule based and machine learning based approaches. It uses the strength of both the approaches while eliminating its weakness. A Hybrid NER called LTH system was proposed by Mikheev which uses a document centered approach. A NER that used the combination of hand-crafted rules, HMM and Maximum Entropy Model was also proposed by Srihari *et al.* (2000).

NER has been designed for scientific, religious text and also emails by Etzioni *et al.* (2005). Lot of research work has been done in NER for biomedical domain by Kazama *et al.* (2002), Califf and Mooney (2003) and Saha *et al.* (2009). NER for agriculture domain called AGNER has been proposed to identify the various crops and pesticide names by Biswas *et al.* (2015). But there is no work that has focused to design NER for travel and tourism domain. Rule Based NER provides results that have high precision whereas the Statistical method provides high recall. Hence, it is efficient and easier to tune statistical system to improve the precision which is better than the effort that is put to tune rule based system to increase the recall.

Tmchem is a very high performance approach that used ensemble machine learning method for NER, designed by Leaman *et al.* (2015). DB pedia data is analyzed for named entity recognition and linking of tweets by Derczynski *et al.* (2015). Biomedical Named Entity Recognition (BNER) which used clustering-based representation, distributional representation andword embeddings for representing word features for NER designed by Tang *et al.* (2014). Yang *et al.* (2015) used

semi-markov's conditional random fields machine learning algorithms to explore the features of two phased bio-medical NER.

Now big data is a buzz word everywhere. Li *et al.* (2015) designed an NER using MapReduce paradigm for biomedical domain. Lao named entity recognition proposed by Yang *et al.* (2015) uses simple heuristic information along with conditional random fields algorithm. Konkol *et al.* (2015) proposed a language independent NER using latent semantics which used unsupervised methods for extracting new features. Neural architecture based NER is designed by Lample *et al.* (2016) which gave great performance without using any language specific knowledge base.

In the current scenario, all applications related to Natural language processing rely on a domain independent NER for their needs with development of domain dependent NER is on the increase in the areas of bioinformatics, agriculture, etc. travel and tourism domain is the most booming and essential industry worldwide. There is more advancement in tourism economics that has enabled us to collect huge amounts of travel and tourism data. If the data are analyzed, it could be a source of high knowledge and intelligence that provides real-time, incremental and right decision making and for the provision of travel tour recommendations. There is no significant research that has been done in NER for Travel and tourism industry and we have decided to go ahead with the CRF based machine learning approach for this NER as the rule based approach cannot handle all possible scenarios and a HMM based approach would require a huge data corpus. As the application for Travel and tourism NER is multifold, we proposed a set of tags as named entities over and above the generic domain NER. These tags are very useful in information retrieval and question answering for tourism domain specific data and very important useful preprocessing tool for tourism data analytics.

MATERIALS AND METHODS

The system design flow is given in Fig. 1. The research is focused on developing a NER for specific travel and tourism domain. The input is the domain specific text and the output is the keywords which are tagged with domain specific defined NER tags.

The travel, tourism, hotel and POI Interest related domain information are collected from the travel websites like trip advisor and wikipedia which is taken as input corpus. These unstructured texts collected from web should be preprocessed and cleaned in order for further processing. The preprocessed input is tagged by Parts of Speech tagger (POS) that help us to identify the context of the keywords such as noun, adverb adjective. The Stanford POS tagger is used which is a generic domain POS tagger. After learning the context of the words it is given as an input to the tokenizer. The tokenizer tokens the given input and stores it in a .csv file which enables us to label the tokens with the appropriate tags. This tokenized labeled input training data set is given as an input to the CRF classification algorithm which based on the labeled input trains and models the travel and tourism domain NER system. When a travel and tourism domain specific document is given as a test data to the CRF algorithm it labels the domain keywords with the trained tags. The detailed description of the proposed design is explained in the section below.

Data collection : There is no standard dataset or corpus for travel and tourism domain. Hence, we manually collected and curated Travel and Tourism domain data from Wikipedia and TripAdvisor.com. These data are used for training and testing purposes. The input corpus consists of text related to tourism, travel, point of interest, hotels, amenities, tourist interest and all words related to domain. The input corpus which is manually curated has 6996 sentences and 140481 words

Data pre-processing: The given input data is preprocessed to remove the irrelevant data. The irrelevant data is the punctuation and stop words. Since the NER is dealing with text, the images are also removed. The data is cleaned and given as an input to the POS tagger.

Generic domain POS tagging: The standard Stanford Generic Domain POS Tagger is used to learn the cleaned input text. It is used to tag each word as noun, adverb, verb, adjective, etc. It can be used to learn the context of the words which is based on relationship with adjacent and related words.

Tokenizer: The preprocessed cleaned data after learning the context is given as input to the Tokenizer. The job of the tokenizer is to split each of the input text tokens and save it in a excel file as .csv file. Tokenizing and saving it as a .csv file is to facilitate preparation of the training data.

Travel and tourism domain vocabulary : For developing NER various keywords that are domain specific are necessary. Travel and tourism domain vocabularies and dictionaries are collected from wordpress.com and Oxford dictionary of travel and tourism which provided the basic and standard keywords for domain. Based on these keywords the tokenized input can be labeled with their respective tags.

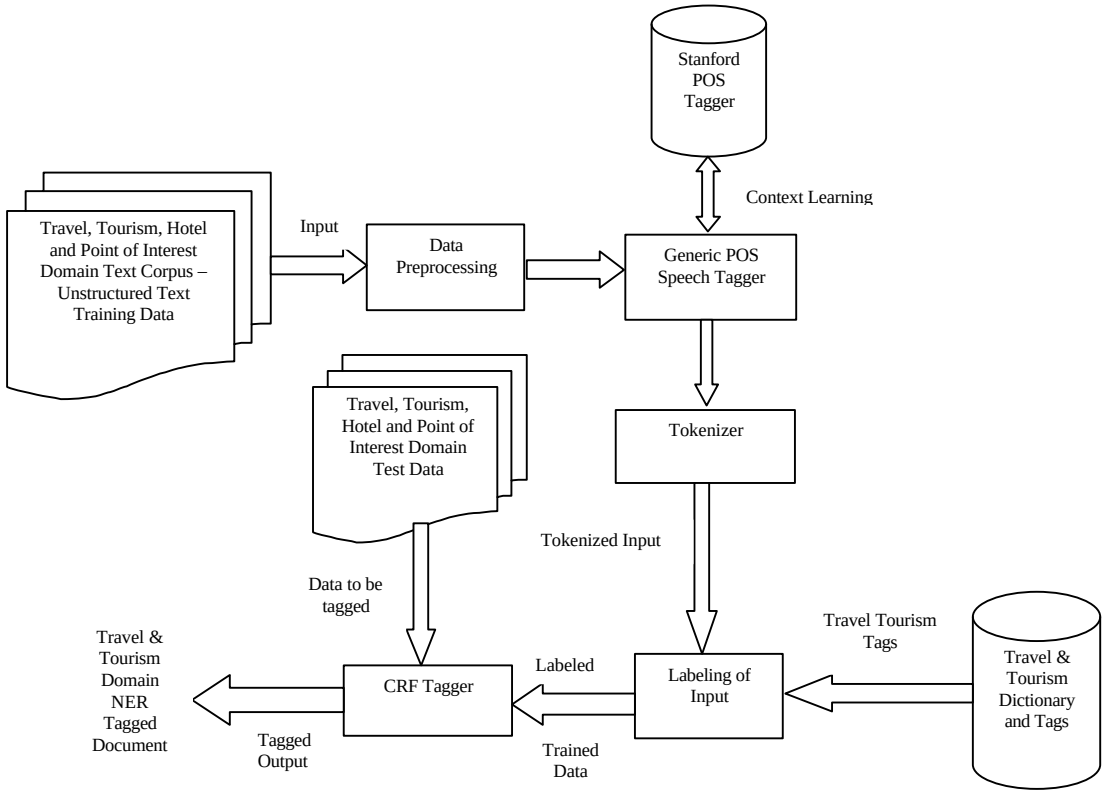

Fig. 1: Travel and tourism domain specific NER architecture

Ner tags for travel and tourism domain: In this research, we considered the travel and tourism vocabulary and various standard ontologies for tourism, hotel and accommodation domain and the various entities/tags that are useful for information extraction are defined as named entities. The details of the entities are listed in Table 1.

Labeling the input: The tourism domain keywords are identified using the dictionary and vocabulary. Semantic and domain specific knowledge is applied on the training corpus and used for tagging the given input data. The newly defined NER tags are also defined. Now each and every tokenized input are labeled with their respective tags and saved. Only the domain keywords are tagged. This serves as the input data for CRF classification algorithm. Since CRF is a semi supervised algorithm the NER for travel and tourism has to be modeled with our training data.

CRF tagger: The preprocessed data need to be trained, for the machine to learn the tags. Each and every preprocessed data is tokenized and input is labeled. Entities related to travel and tourism domains are labeled with their new tags. The non-domain entities are labeled as 'O'. BIO2 Annotation standard is used. The Beginning and intermediate positions of the entities are marked with B-and I-tags. The training set consists of words that are tokenized and labelled with appropriate tags and where the test set consists of words and sentences.

The statistical modeling method used to learn the labeled test data is Conditional Random Fields (CRFs). It is a machine learning algorithm that is widely used for pattern recognition and structured prediction. . CRF is a discriminative undirected probabilistic graphical model. CRF algorithm is the most efficient than any other machine learning algorithm because it predicts sequence of labels for sequence of input text taking the context of the input into account. It is a semi-supervised learning algorithm that can be best used for Named entity recognition of travel and tourism domain according to Liao and Veeramachaneni (2009).

The given input data and its labeled NER is learnt using CRF algorithm which was used by Jenny and colleagues. The test data set used is again related to travel and tourism domain. The CRF algorithm based on the learned input tags the domain related keywords and entities of the test data.

Test data : The test is data also collected from the tourism websites such as TripAdvisor.com and Wikipedia which consists of information about various places. The test data document is given as input to the CRF classifier. The classifier based on the trained input data labels the keywords of the document with the appropriate tags.

Output : Output is the tagged travel and tourism domain keywords which could be used for further processing for travel and tourism data analytics.

RESULTS AND DISCUSSION

According to Wikipedia, the evaluation of the NER task is done by Precision (P) and Recall (R) and F-Measure. Precisionis defined as the percentage of entity labeled is correct with respect to thegold standardevaluation data. In the area of Information Retrieval the fraction of documents that are retrieved which are relevant to the query is called as Precision.

$$\text{Precision} = \frac{|\{\text{No. of Tourismdomain key words in the document}\} \cup \{\text{No. of Tourismdomain keywords tagged}\}|}{|\{\text{No.of Tourism domain keywords tagged}\}|} \quad (1)$$

Precision in other words is the number of true positives which is the number of words correctly tagged as travel and tourism domain keywords divided by the sum of the true positives and false positives which is number of the words tagged as domain keywords by our Travel and Tourism domain NER.

Recall is the measures of the number of names in the gold standard that are present at exactly the same location in the predictions are correctly labeled. Precisely in the field of Information Retrieval the fraction of documents which are relevant to that has been successfully retrieved is called as Recall.

$$\text{Precision} = \frac{|\{\text{No.of Tourism domain key words in the document}\} \cup \{\text{No.of Tourism domain keywords tagged}\}|}{|\{\text{No.of Tourism domain keywords tagged}\}|}$$

Recall in other words is the number of true positives which is the number of words correctly tagged as travel and tourism domain keywords divided by the sum of the true positives and false negatives which is the actual number of relevant words related to travel and tourism domain.

Precision and recall result are combined together to form F-measure of NER performance. It is calculated by the uniformly weighted harmonic mean of precision and recall.

$$F = 2.\frac{\text{precision.recall}}{\text{precision} + \text{recall}}$$

Accuracy of our named entity recognizer can be determined by the sum of the keywords correctly tagged as travel domain and non-domain keywords by the total number of words in the document.

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

Where:

tp = The true positive value-number of words correctly tagged by our NER

tn = The true negative value-number of words correctly classified as non-tourism domain keywords

fp = The false positive value-number of words wrongly tagged as travel and tourism domain keywords by our NER and 'fn' is the false negative value-number of travel and tourism domain keywords which is not tagger by our NER

We have made an evaluation for travel and tourism NER using domain related input corpus collected from Wikipedia and from World Wide Web. The system shows more accuracy for travel related data compared to the Stanford NER Too.l

Table 2 Evaluation with the performance of our domain specific NER for domain specific output. We compared our output with the travel and tourism dictionary to find out the accuracy of our NER. Out of tagged 71645 words, 59465 words are correctly tagged as travel and tourism domain keywords. Thus our travel and tourism domain has a precision of 85%. 12180 words are wrongly classified as travel and tourism domain keywords. 13053 keywords were not tagged as travel domain keywords by our NER System. So our system had a recall of 82%. The F measure which is the harmonic mean of precision and recall is 83%. Accuracy of our NER system is 82%.

Table 1: Domain specific tags for travel and tourism domain

Travel and tourism domain tags	Explanation	Need for travel and tourism tags
Hotel name	Name of the hotel	Identifying hotel name enables to find the reviews and related features of the hotel
POI	Point of Interest to tourist in various locations	Identifying point of interest enables to find the interesting location of tourist interest
POI-Type	Nature of point of interest like sights and Landmarks, Outdoor Activities, Nature and Parks, Fun and Games, Shopping, Museum, Tour and Activity, Museums. Amusement Parks, Theatre and Concert, Zoo and Aquariums, Spa and Wellness	the nature of point of interest helps us to find the characteristics of the place which further enables fast query processing and information retrieval according to traveler interest
Amenities	The facilities provided by Hotels and Point of Interest like Swimming Pool, Laundry, Free Breakfast etc	Identifying amenities of the hotel helps us to identify easily what facilities the hotel has and retrieve according to traveler's priority
Star rating	Rating of Hotel such as 5 star deluxe, 4 star, 3 star, 2 star	identifying star rating of the hotels help us to classify hotel according to their rating
Heritage hotel	Heritage, heritage classic, heritage grand	Helps us to identify the heritage hotels and classify it
Transportation	bus-stop, airport, railway station	identifying the nearest bus-stop, airport and railway station to the location will be helpful for the traveler to move on
Review	includes positive and negative review	Analyzing and tagging the positive and negative keywords help in further processing during opinion mining
Hotel style	Best value, boutique, budget, business, charming, classic, family-friendly, luxury, mid-range, quaint, quiet, resort hotel, romantic, trendy	These tags help us to identify the type of hotel according to the taste and requirement of various people
POI-characteristic	nature, scenic, pleasant, romantic	These tags helps us to spot the characteristic of the places of interest and make it easy to choose the destination
Restaurant name	Name of the restaurants	These tags helps in finding the names of the restaurant in the particular locality. It will be distinguished form name of the person
Cuisine type	Chinese, Indian, Tandoori, Italian, Mexican, Arabian, Barbeque	The types of cuisine of a restaurants can be defined by this tags
Restaurant type drinks	Dessert, Coffee and Tea, Bakery, Bar and Pub	Helps to identify and classify the type of restaurant based on type of food and available
Tourism type	Natural beauty, medical tourism, heritage sites, wildlife, ayurveda and wellness, royal retreats, spirituality, adventure sports, cool retreats, desserts, beach, weekend getaways	Helps us to identify the type of tourism the traveler interested to go and visit
Climate	Hot, Cold, Rainy	Tags the climatic condition in various months
Season	Summer, Winter, Autumn, Spring	It helps in identifying the types of seasons present in various point of interest and appropriate seasons to visit the place.
Longitude	The longitude of a particular place	Pinpoints the longitude of a place to identify the exact location
Latitude	The latitude of a particular place	Pinpoints the latitude of a place to identify the exact location
Pin code	Pin code of region	Identify the pin code of the location. Pin code is a very unique number to identify a place
Opening time	Opening Time of restaurants, point of interest	This tag is important to identify the opening time of the restaurant and point of interest
Closing time	Closing time of restaurants, point of interest	It identifies the closing time of the restaurant and points of interest
On season	Apt month to visit the place	Tags suitable month to visit a place or point of interest
Off season	Off Seasonal months	Tags the off seasonal month of a place
Specialty	Specialty of particular place	Identifies the specialty or interest of various places and restaurants
Distance	The number preceding the km	Identifies the number which specifies the distance

Table 2: Evaluation for domain specific input

System	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
Travel and Tourism Domain NER	85	82	83	82

The input document consists of 6996 sentences and 140481 words which are specific to Travel and Tourism Domain. In order to evaluate our NER the same domain related document is given as an input to both Stanford NER and Travel & Tourism domain NER, in order to compare the performance.

The 7 class Stanford's named entity tagger tags only 22% of the words into various entities such as location, person, organization, money, percent, date and time. By using our domain specific NER, the tags increased by 29%. So, there is an increase in the number of tags due to entities related to travel and tourism domain. Our NER system was not able to classify the names of the Point of Interest in most of the cases. Similarly our NER system was not able to define the names of the hotels and restaurants due to which there was decrease in our precision and recall. Our ner has tagged 59465 out of 72518 words of the tourism domain keywords as true positive which gives correctness of 82%. 55783 words non domain keywords are correctly classified as true negative.

Table 3: Confusion matrix

	Predicted Condition			
Total population 140481 words	Positive	Negative	Prevalence 52 (%)	Prevalence
True condition				
Conditionpositive 72518 words	True Positive 59465 words	False Negative 13053 words	Positive Rate (TPR) Sensitivity, Recal l82 (%)	False Negative Rate (FNR), Miss Rate18 (%)
Condition negative 67963 words	False Positive words12180	True Negative 55783 words	False Positive Rate (FPR), Fallout 18 (%)	True Negative Rate (TNR), Specificity82%
Accuracy 82 (%)	Positive Predicted Value (PPV), Precision 85(%)	False omission rate 19 (%)	Positive likelihood ratio 4.56	Diagnostic odds ratio 20.72
False discovery Rate 17 (%)	Negative predictive rate 81 (%)	Negative likelihood ratio 0.22		

Table 4: Analysis of tags

No. of Input Words	No. of keywords tagged stanford NER Tagger	No. of keywords tagged using travel and tourism domain NER tagger	No. of non-domain keywords	No of new tags	Old tags replaced
140481	30906	71645	37930	25	Names Replaced As hotel and restaurant nametime replaced as opening and closing timeNumbers replaced with distance and pin codesorganization tags replaced as POI-TypeMonths are tagged as Off season and on season

A confusion matrix is used to enable us to know the performance of the classification algorithm to facilitate in defining tags for our Travel and Tourism Domain Area. Table 3. helps us in analyzing the performance of our system using various measures using confusion matrix The standard tag which is named as NAME is replaced with name of the hotels, restaurants and point of interest. The time can be classified as opening time and closing time related to various point of interest. Table 4 analyzes the number of input tags tagged using Stanford NER and our Travel and Tourism Domain specific NER.

We need to improve our NER system by providing various training set to include all possible classification related to Travel and Tourism to enable the machine to learn better. But our system proved to be more effective for documents related to travel and Tourism domain than our Generic Domain Stanford's NER tagger. The Standard Stanford NER tagger was able to tag only 30906 keywords from the document which are all Generic Domain entities. Thus our NER system proved best to identify the domain related keywords.

CONCLUSION

In this research, we have developed a named entity tagger for travel and tourism domain. In this work we took travel and tourism domain input corpus as an input. After various preprocessing steps, the tokenized data is labeled with the new tags defined. The input labeled data is given as an input to CRF classification algorithm which learned and builds a suitable NER model for travel and tourism domain. When a domain related input is given as test data it tags the keywords of the document with the appropriate tags.

Due to unavailability of standard travel and tourism data we cannot state the accuracy but it can be improved in the future keeping this named entity tagger as a base. It can be used as a data preprocessing step to collect entities of a travel and tourism domain. It will be helpful for researchers to design a structured dataset based on the unstructured input which can be used for further analytics for the development of the travel and tourism domain. This research work will facilitate the research work in the travel and tourism domain for mining and recommendations.

REFERENCES

Bikel, D.M., R. Schwartz and R.M. Weischedel, 1999. An algorithm that learns whats in a name. Mach. Learn., 34: 211-231.

Biswas, P., A. Sharan and A. Kumar, 2015. AGNER: Entity tagger in agriculture domain. Proceedings of the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), March 11-13, 2015, IEEE, New Delhi, India, ISBN: 978-9-3805-4415-1, pp: 1134-1138.

Borthwick, A., 1999. A maximum entropy approach to named entity recognition. Ph.D Thesis, New York University, New York, USA. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.641.2654&rep=rep1&type=pdf

Califf, M.E. and R.J. Mooney, 2003. Bottom-up relational learning of pattern matching rules for information extraction. J. Mach. Learn. Res., 4: 177-210.

Carreras, X., L. Marquez and L. Padro, 2002. Named entity extraction using adaboost. Proc. Conf. Nat. Lang. Learn., 20: 1-4.

Coates, S.S., 1992. The analysis and acquisition of proper names for the understanding of free text. Comput. Humanities, 26: 441-456.

Curran, J.R. and S. Clark, 2003. Language independent NER using a maximum entropy tagger. Proceedings of the Seventh Conference on Natural Language Learning, May 31-June 1, 2003, ACM, Stroudsburg, Pennsylvania, USA., pp: 164-167.

Derczynski, L., D. Maynard, G. Rizzo, V.M. Erp and G. Gorrell et al., 2015. Analysis of named entity recognition and linking for tweets. Inf. Process. Manage., 51: 32-49.

Etzioni, O., M. Cafarella, D. Downey, A.M. Popescu and T. Shaked et al., 2005. Unsupervised named-entity extraction from the web: An experimental study. Artif. Intell., 165: 91-134.

Finkel, J.R., T. Grenager and C. Manning, 2005. Incorporating non-local information into information extraction systems by gibbs sampling. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, June 25-30, 2005, ACM, Stroudsburg, Pennsylvania, USA, pp: 363-370.

Fog Computing World, 2014. Internet of things to generate zettabytes of data by 2018. Fog Computing Publishing, Palo Alto, California. http://www.fogcomputingworld.com/topics/fog-computing/articles/393474-internet-things-generate-zettabytes-data-2018.htm

Freitag, D. and N. Kushmerick, 2000. Boosted wrapper induction. Proceedings of the 17th National Conference on Artificial Intelligence, July 31-August 2, 2000, AAAI Press/The MIT Press, Austin, TX., pp: 577-583.

Ge, Y., H. Xiong, A. Tuzhilin and Q. Liu, 2014. Cost-aware collaborative filtering for travel tour recommendations. ACM. Trans. Inf. Syst. TOIS., Vol. 32, 10.1145/2559169

Kazama, J.I., T. Makino, Y. Ohta and J.I. Tsujii, 2002. Tuning support vector machines for biomedical named entity recognition. Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain, July 11-12, 2002, ACM, Stroudsburg, Pennsylvania, USA, pp: 1-8.

Konkol, M., T. Brychcin and M. Konopik, 2015. Latent semantics in named entity recognition. Expert Syst. Appl., 42: 3470-3479.

Lafferty, J.D., A. McCallum and F.C.N. Pereira, 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, June 28-July 1, 2001, Williamstown, MA., USA., pp: 282-289.

Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, 2016. Neural architectures for named entity recognition. Procceedings of the International Conference on NAACL, April 7-7, 2016, Cornell University Library Press, New York, USA., -.

Leaman, R., C.H. Wei and Z. Lu, 2015. TmChem: A high performance approach for chemical named entity recognition and normalization. J. Cheminf., 7: S1-S3.

Lee, K.J., Y.S. Hwang and H.C. Rim, 2003. Two-phase biomedical NE recognition based on SVMs. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, June 29-July 3, 2003, ACM, Stroudsburg, Pennsylvania, USA, pp: 33-40.

Li, K., W. Ai, Z. Tang, F. Zhang and L. Jiang et al., 2015. Hadoop recognition of biomedical named entity using conditional random fields. IEEE. Trans. Parallel Distrib. Syst., 26: 3040-3051.

Liao, W. and S. Veeramachaneni, 2009. A simple semi-supervised algorithm for named entity recognition. Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, June 4-4, 2009, ACM, Stroudsburg, Pennsylvania, USA, ISBN: 978-1-932432-38-1, pp: 58-65.McNamee, P. and J. Mayfield, 2002. Entity extraction without language-specific resources. Proceedings of the 6th Conference on Natural Language Learning, August 31-September 1, 2002, ACM, Stroudsburg, Pennsylvania, USA, pp: 1-4.

Minkov, E., R.C. Wang and W.W. Cohen, 2005. Extracting personal names from email: Applying named entity recognition to informal text. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, October 6-8, 2005, ACM, Stroudsburg, Pennsylvania, USA., pp: 443-450.

Munro, R. and C.D. Manning, 2012. Accurate unsupervised joint named-entity extraction from unaligned parallel text. Proceedings of the 4th Workshop on Named Entity, July 12-12, 2012, ACM, Stroudsburg, Pennsylvania, USA, pp: 21-29.

Rau, L.F., 1991. Extracting company names from text. Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications, February 24-28, 1991, Miami Beach, FL., pp: 29-32.

Saha, S.K., S. Sarkar and P. Mitra, 2009. Feature selection techniques for maximum entropy based biomedical named entity recognition. J. Biomed. Inf., 42: 905-911.

Soderland, S., 1999. Learning information extraction rules for semi-structured and free text. J. Mach. Learn., 34: 233-272.

Srihari, R., C. Niu and W. Li, 2000. A hybrid approach for named entity and sub-type tagging. Proceedings of the Sixth Conference on Applied Natural Language Processing, April 29-May 4, 2000, ACM, Stroudsburg, Pennsylvania, USA, pp: 247-254.

Tang, B., H. Cao, X. Wang, Q. Chen and H. Xu, 2014. Evaluating word representation features in biomedical named entity recognition tasks. BioMed. Res. Int., Vol. 2014,

Thielen, C., 1995. An approach to proper name tagging for German. Procceedings of the Conference on European Chapter of the Association for Computational Linguistics, June 28-29, 1995, SIGDAT Publisher, University of Tubingen Publisher, Tubingen, Germany, pp: 1-7.

World Travel & Tourism Council, 2015. Economic Impact of Travel and Tourism. World Travel & Tourism Council, London, England. https://www.wttc.org/research/economic-research/economic-impact-analysis/

Yang, M., L. Zhou, Z. Yu, S. Gao and J. Guo, 2015. Lao named entity recognition based on conditional random fields with simple heuristic information. Proceedings of the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), August 15-17, 2015, IEEE, Kunming, China, ISBN: 978-1-4673-7682-2, pp: 1426-1431.