ISSN:0975-9646

Survey on Name Entity Recognition Used Machine Learning Algorithm

Daljit Kaur
PTURC
SSIET, Derabassi, Punjab

Ashish Verma
Asst. Professor ,Deptt. of CSE & IT
SSIET, Dera Bassi, Punjab

Abstract*—The amount of textual information available electronically has made it difficult for many users to find and access the right information within acceptable time. Research communities in the natural language processing (NLP) field are developing tools and techniques to alleviate these problems and help users in exploiting these vast resources. These techniques include Information Retrieval (IR) and Information Extraction (IE). The work described in this thesis concerns IE and more specifically, named entity extraction in English. The English language is of significant interest to the NLP community mainly due to its political and economic significance, but also due to its interesting characteristics. Text usually contains all kinds of names such as person names, company names, city and country names, sports teams, chemicals and lots of other names from specific domains. These names are called Named Entities (NE) and Named Entity Recognition (NER), one of the main tasks of IE systems, seeks to locate and classify automatically these names into predefined categories. NER systems are developed for different applications and can be beneficial to other information management technologies as it can be built over an IR system or can be used as the base module of a Data Mining application. In this thesis we propose an efficient and effective framework for extracting Arabic NEs from text using a rule based approach. Our approach makes use of English contextual and morphological information to extract named entities. The context is represented by means of words that are used as clues for each named entity type. Morphological information is used to detect the part of speech of each word given to the morphological analyzer. Subsequently we developed and implemented our rules in order to recognize each position of the named entity. Finally, our system implementation, evaluation metrics and experimental results are presented.

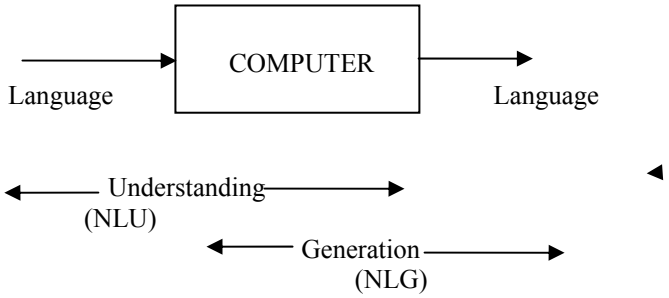
***Index Terms*—Natural Language Processing, Machine Translation, Name Entity Recognition, Different Languages.**

I. INTRODUCTION

Natural Language Processing (NLP) is the field whose aims to convert the human language into the formal representation that is easy to manipulate for the computer. The application includes, information retrieval, information extraction, speech recognition summarization, machine translation, search and computer interfaces for human.

Natural Language Processing (NLP) has various Tasks:-

1) Part-of-Speech Tagging (POS):- labelling each word with a unique tag which indicates its syntactic role, e.g. verbs, noun, adverbs, adjectives.
2) Chunking: - aims at labelling segments of a sentence with syntactic constituents such as verb or noun phrase (VP or NP). Each word is assigned only one unique tag, often encoded as a begin-chunk or inside-chunk tag.
3) Named Entity Recognition (NER):-labels elements in the sentence into categories viz "person", "company", or "location".
4) Semantic Role Labelling (SRL):- aims at giving a semantic Role to a syntactic constituent of a sentence. The precise arguments depend on a verb's frame and if there are multiple verbs in a sentence some words might have multiple tags. In addition to the ARG0-5 tags, there are 13 modifier tags such as ARGM-LOC (locational) and ARGM-TMP (temporal) that operate in a similar way for all verbs.
5) Languages Models:-A language model traditionally which estimates the probability of the next word being "w" in a given sentence.
6) Semantically Related Words ("Synonyms"):- It is the task of predicting whether two words are semantically related which is measured using the Word Net database as ground truth [1].



II. NER AND ITS APPROACHES

NAME ENTITY RECOGNITION AND ITS APPROACHES

Name Entity Recognition:- Named Entity Recognition is the process of identification and classification of all proper nouns in a given text document or a sentence into pre-defined classes such as persons, locations, organizations, date, address and time expressions. Named Entities are defined as the proper names identified in a text. Identified text may be a person's names, organization's names, location's names, and date and time expressions. To make a computer acceptable and divide these named entities into pre-defined categories, which are important tasks of NLP. This task is defined as Named Entity Recognition. It is also called Information Extraction [15].

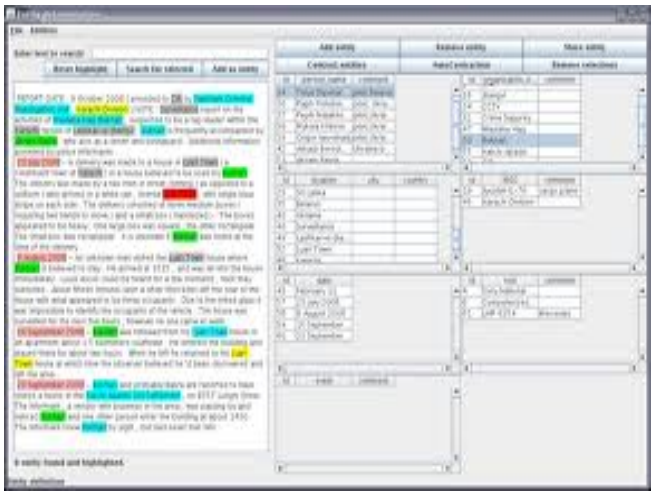
For example:-

Name entity type	**Examples**
ORGANIZATION	Global India
PERSON	President Pranab Mukherjee, Navneet
LOCATION	Chandigarh, Mount Everest
TIME	three fifty a m, 12:30 p.m.

MONEY	$567,175 million Canadian Dollars
DATE	12-06-1991, June
PERCENT	25.22 %, fifty pct,
FACILITY	Stonehenge Washington
GPE	Greenland, Bhutan

APPLICATIONS OF NAMED ENTITY RECOGNITION

Name entity recognition is useful in many Natural Language Processing applications, like- information retrieval, extraction information, question answering (true or false), parsing, and machine translation (from one language to another), the metadata for the Semantic.NER can also give the information to users who are looking for person or organization names with fast information [8] .Name entity recognition systems are used in the areas of entity identification in the field of medical images. Earlier, NER systems were used by primarily extraction from journalistic articles and then Automatic Content Extraction (ACE) evaluation also included several types of text styles, such as WebPages and detects text from the speech or any audio.



APPROACHES FOR NAME ENTITY RECOGNITION

There are different approaches to name entity recognition. It can be categorized into two broad categories:-

A) Rule based (Linguistic) approaches: - Rule based approaches rely on hand-crafted rules, written by language experts, to recognize and classify NEs. Rule-based approaches may contain Lexicalized grammar, Gazetteer lists, List of triggered words etc. [17]. There are two disadvantages for using this approach: first is to developing and maintaining rules and dictionaries is a tedious and costly task. Second these systems cannot be transferred to other languages or domains.

B) Machine learning (Statistical) approaches:- Machine learning approaches rely on statistical models to make predictions about name entities in given text. Large amounts of annotated training data are required for these models to be effective, which can prove costly [8]. There are three main machine learning approaches:-Supervised, Semi-supervised, Unsupervised.

a) **Supervised Learning:-** Supervised learning approaches build predictive models based on the labelled data and true labels. Some of the supervised machine learning techniques is:

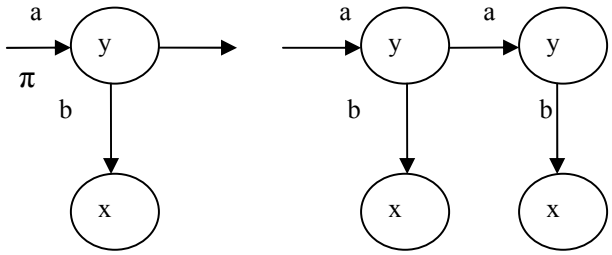
- Hidden Markov Model (HMM)
- Decision Trees
- Maximum Entropy (MaxEnt)
- Support Vector Machines (SVM)
- Conditional Random Fields (CRFs)

I. HIDDEN MARKOV MODEL (HMM)

HMM is a probabilistic automata based on markov model where a label corresponds to a state and an observation symbol to a word at a state. Both state transition and observation symbols are described in probabilistic manner. HMM has a model M=(O,Q,A,B,π) where

$$A=\{a_{ij}, i, j=1.....N\},\ B=\{b_i(y_t), i=1,.....t=1\}$$

and O ,Q mean a finite set of observation symbol of x and y.[17]



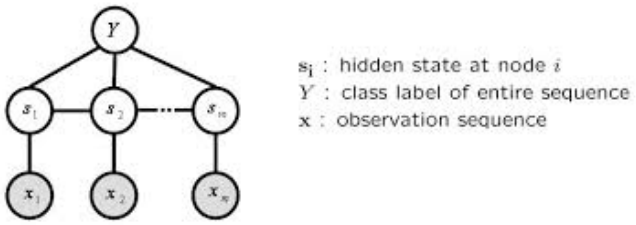
II. MAXIMUM ENTROPY (MAXENT)

ME conditional probabilistic sequence model. In this multiple features are extracted from one word and handle their dependency for the long term. Maximum entropy is that in which model for least biased that considers all known facts is the one which maximizes entropy. Every source has a model of exponential that takes the inspection feature as input and distribution over possible next state as a output. The result output labels are associated with states. It solves the problem of multiple feature representation and long term dependency issue occurred in HMM. It has increased the recall and greater precision than Hidden Markov Model. The probability conversion leaving any given state must sum to one, so, it is influence towards that states with lower or less outgoing transitions. The state with one outgoing state transition will ignore all observations. To overcome Label Bias Problem we can change the state-transition structure or we can start with fully connected model and let the training procedure decide a good structure [19].

III. CONDITIONAL RANDOM FIELDS (CRF)

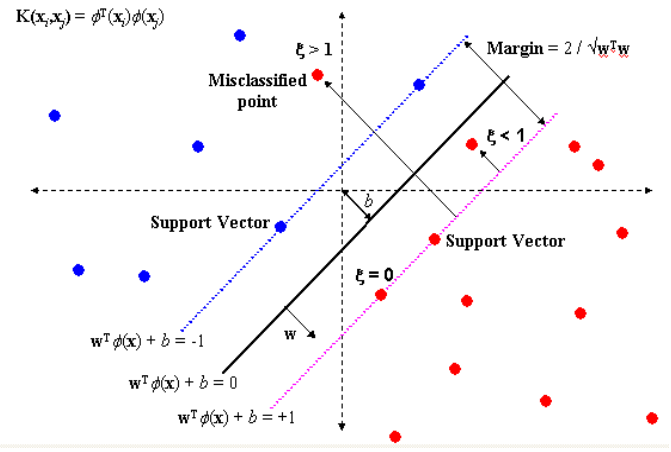
CRF is a type of discriminative probability model. It has all the advantage of maximum entropy instead the label bias problem. CRFs are undirected graphical models and also called random fields which are used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes.

Random field:- Let G = (Y, E) be a graph where each vertex YV is a random variable. Suppose P(Yv-all other Y) = P(Yv-neighbours(Yv)), then Y is a random field. Let X = random variable over data sequences to be labelled Y = random variable over corresponding label sequence. Definition Let G = (V, E) be a graph such that Y = (Yv) vV, so that Y is indexed by the vertices of G. Then (X,Y) is a conditional random field , when conditioned on X, the random variables Yv obey the Markov Property with respect to the graph: P(Yv —X,Yw, w v) = P(Yv —X,Yw, w v), where 'w' 'v' means that 'w' and 'v' are neighbours in G.



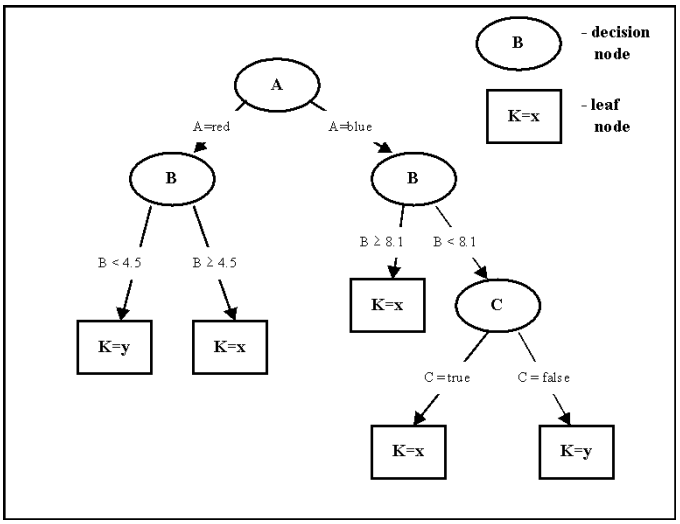
IV. SUPPORT VECTOR MACHINES (SVM)

The SVM is based on discriminative approach which is use for positive and negative examples to learn the variance between the two classes. The SVMs are known to robustly manage large feature sets and to develop models that maximize their generalizability Take two set of training data for a two-class problem: {(*x*1, *y*1),..... (*xN*, *yN*)}, where xi ε RD is a feature vector of training data of ith sample and yi ε {+1, -1} is the class to which xi related. The main goal is to find a decision function that accurately predicts class y for an input vector x. A non-linear support vector machine classifier states a decision function f (x) = sign (g (x)) assumed value of f(x) is 1for this section. Here, f(x) =+1means x is a member and f(x) =-1 means x is not a member of a certain class. zi called support vector and representative of training examples, m is the number of support vectors. So the computational complexity of g(x) is proportional to m. SVM and different constants are determined by solving a certain quadratic programming problem, K(*x*, *zi*) is a kernel that implicitly maps vectors into a higher dimensional space. Typical kernels use dot products (*K*(*x*, *zi*) = *k*(*x*.*z*)) .A polynomial kernel of degree d is given by *K*(*x*, *zi*) = (1+x)d It can use different kernels, and the design of each kernel for a particular application is an important research issue [21].



V. DECISION TREE

Decision Tree is a popular and powerful tool for categorizing and forecast. Rules are used for artificial intelligence and neural network in decision tree. That rules can easily be expressed so that human can well understand and directly use rules in a database access language like SQL so that records failing into a particular classification may be tree. Decision Tree is a classifier in the form of a tree structure where each node represent a leaf node, indicates the value of the output attributes of expressions, a decision, specifies some text to be carried out on a single attribute value with one branch and sub-tree for each possible outcome of the text. It is an inductive approach to acquire knowledge on classification [22].

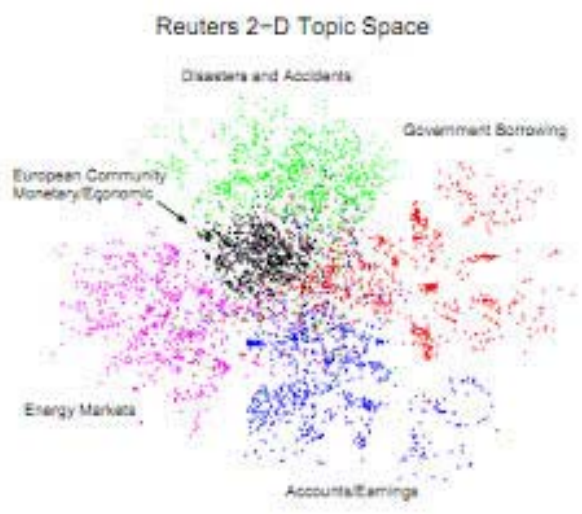


b) Unsupervised Learning

Unsupervised learning approaches don't expect any implicit or structural information about the data they are processing. The typical approach to unsupervised learning is clustering. For example, one can try to collect names from clustered groups based on the similarity of context. There are other methods also, which are unattended. Basically, the techniques based on lexical resources (e.g. WorldNet) calculated on lexical patterns and statistics on a large unannotated corpus.

c) Semi supervised Learning

The term semi-supervision or weak supervision is still relatively young. The main SSL technology is called bootstrapping and includes a small measure of control, like a row of seeds, for the beginning of the learning process. In semi-supervised approach, a model is trained on an initial set of labelled data and true labels, then, predictions are made on a separate set of unlabeled data, and then improved models are created iteratively using predictions of previously developed models. [16]. For example, a system aimed at "disease names" could prompt the user to give a small number of example names.



VI. RELATED WORK

Many researchers have been discussed about Name entity recognition of machine translation.

Deepti Bhalla [23] in this name entity comprises two tasks; they can be translated or transliterated with the help syllabification. In this translation of English to Punjabi by using statistical rule based approach. Syllabification algorithm is used for translation of name entity. They calculated n-gram probability for syllable.

Kamal deep [24] rule based approach is used for addressed the problem of transliterating Punjabi to English language. The proposed transliteration scheme uses grapheme based method to the transliteration problem.

Sharma et al. [5] show English-Hindi transliteration by using statistical machine translation in the different notation. This paper WX-notation gives the better result over UTF –notation by English Hindi corpus by using phrase based statistical machine translation.

Dhore et al [7] have addressed the problem of MT where give named entity in Hindi using Devanagari script by using conditional random field as a statistical probability tool. In this approach, they show machine transliteration of name entities for Hindi-English language using CRF as statistical probability tool. The accuracy of this system is 85.79%.

Sweta Kulkarni [15] this paper shows the survey of name entity recognition. Then, they describe the various approaches used for name entity recognition, followed by the Performance Metrics which is used to evaluate the system of name entity. They consider the existing NER systems for each of the four main South Indian languages: Kannada, Telugu, Tamil, and Malayalam and analyze them.

Nusrat Jahan [17] in this paper they describe the various approaches used for NER and summery on existing work done in different Indian Languages using different approaches and also describe introduction about Hidden Markov Model (HMM) and the Gazetteer method for name entity recognition. We also present some experimental result using Gazetteer method and HMM method that is a hybrid approach. Finally in the last the paper also describes the comparison between these two methods separately and then we combine these two methods so that performance of the system is increased.

Ryohei Ageishi [18] combination of statistical with rule based approach is used to recognize name entity in the morphological analysis. HMM is use for tagging the English text. They discuss rule based approach over n consecutive word for the rule extraction.

Thoudam Doren Singh [21] there are two different models, one using an active learning technique based on the context patterns generated from an unlabeled news corpus and the other based on the well known Support Vector Machine have been developed. The Manipuri news corpus has been manually annotated with the major name entity tags, namely name of the person, Location name, and name of Organization and to apply SVM. The SVM based system makes use of the different contextual information of the words along with the variety of orthographic word-level features which are helpful in predicting the NE classes.

Georgios Paliouras [22] a NERC system assigns semantic tags to phrases that correspond to named entities, such that persons, locations and organisations. Typically, such a system makes use of two language resources: a recognition grammar and a lexicon of known names, classified by the corresponding named-entity types.we evaluated the behaviour of C4.5 on the task of learning decision trees to recognise and classify named entities in text. This approach reduces significantly the effort needed for customising a NERC system to a particular domain.

Yunita Sari [25] in this, to extract important facts from unstructured text which later help to populate database entries. Name Entity Recognition is one of the main task needed to develop text mining systems in which it is used to identify and classify entities in the text into predefined categories such as the person's name, organization's name, locations, dates, times, quantities, percentages, etc. Mainly they focuses on studying the optimum solution to perform name entity recognition. Many algorithms have been reported for NER ranging from simple statistical methods to advanced Natural language processing methods. This paper describes the possibility to apply Link Grammar (LG) and Basilisk Algorithm in NER.

CONCLUSION

NER has been an active research sub-field of AI from years. But the challenges faced during translation need to be solved for which more detailed study of various natural languages is required. So still a lot of work is required to develop a completely automatic translation system. Improved Name entity recognition is most important part of machine translation. There are some characters exist in English which are double meaning like you is also written in u. The major inaccuracies in the transliteration are due to poor word selection. In this paper, there have described the recognition system build on statistical techniques. There are many issues left for further improvement. The system itself could be improved. In this investigation, we have discussed how to recognize name entity.

REFERENCES

[1] Ronan Collobert, Jason Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning"

[2] Harjinder Kaur, Dr. Vijay Laxmi, "A Web Based English to Punjabi MT System for News Headlines," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[3] Latha R. Nair, David Peter S., "Machine Translation Systems for Indian Languages," International Journal of Computer Applications (0975 – 8887) Volume 39– No.1, February 2012.

[4] Vishal Gupta, Gurpreet Singh Lehal, "Named Entity Recognition for Punjabi Language Text Summarization," International Journal of Computer Applications (0975 – 8887) Volume 33– No.3, November 2011.

[5] Shubhangi Sharma, Neha Bora and Mitali Halder, " English-Hindi Transliteration using Statistical Machine Translation in different Notation ," International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.

[6] Rejwanul Haque, Sandipan Dandapat, Ankit Kumar Srivastava, Sudip Kumar Naskar and Andy Way, "English—Hindi Transliteration Using Context Informed PB-SMT: the DCU System for NEWS 2009," CNGL, School of Computing Dublin City University, Dublin 9, Ireland.

[7] Yuxiang Jia, Danqing Zhu, Shiwen Yu, "A Noisy Channel Model for Grapheme-based Machine Transliteration," Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 88–91, Suntec, Singapore, 7 August 2009. c 2009 ACL and AFNLP.

[8] Kamal Deep, Dr.Vishal Goyal, "Hybrid Approach for Punjabi to English Transliteration System," International Journal of Computer Applications (0975 – 8887) Volume 28– No.1, August 2011.

[9] Mitali Halder, Anant Dev Tyagi, "English-Hindi Transliteration by applying finite rules to data before training using Statistical Machine Translation," 978-1-4799-2845-3/13/$31.00 ©2013 IEEE.

[10] Deepti Bhalla, Nisheeth joshi, Iti mathur, "Improving the quality of machine translation output using novel name entity translation scheme," 987-1-4673-7/13/$31.00©2013 IEEE.

[11] Darvinder kaur, Vishal Gupta, "A survey of Named Entity Recognition in English and other Indian Languages," IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010.

[12] Kamaljeet Kaur Batra and G S Lehal, "Rule Based Machine Translation of Noun Phrases from Punjabi to English," IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.

[13] Jiyou Jia, "The Generation of Textual Entailment with NLML in an Intelligent Dialogue System for Language Learning CSIEC," 978-1-4244-2780-2/08/$25.00 ©2008 IEEE.

[14] Harjinder Kaur, Dr. Vijay Laxmi, " a survey of machine translation approaches," international journal of science, engineering and technology research (ijsetr) volume 2, issue 3, march 2013.

[15] Malarkodi, C S., Pattabhi, RK Rao and Sobha, Lalitha Devi, "Tamil NER – Coping with Real Time Challenges", Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 23–38, COLING 2012, Mumbai, December 2012.

[16] Yunita Sari, M. Fadzil Hassan, Norshuhani Zamin, "A Hybrid Approach to Semi-Supervised Named Entity Recognition in Health, Safety and Environment Reports,"International Conference on Future Computer and Communication © 2009 IEEE.

[17] Nusrat Jahan, Sudha Morwal and Deepti Chopra, "Named Entity Recognition in Indian Languages Using Gazetteer Method and Hidden Markov Model: A Hybrid Approach," International Journal of Computer Science & Engineering Technology (IJCSET) ISSN : 2229-3345 Vol. 3 No. 12 Dec 2012.

[18] Ryohei Ageishi, Takao Miura, "Name entity recognition based on a hidden markov model in part of speech tagging," 978-1-4244-2624-9/08/$25.00 ©2008 IEEE

[19]https://www.google.co.in/#q=LANGUAGE+INDEPENDENT+NAMED+ENTITY+RECOGNITION.

[20] Brahmaleen K. Sidhu,Arjan Singhand Vishal Goyal, " Identification of Proverbs in Hindi Text Corpus and their Translation into Punjabi,"JOURNAL OF COMPUTER SCIENCE AND ENGINEERING, VOLUME 2, ISSUE 1, JULY 2010.

[21] Thoudam Doren Singh, Kishorjit Nongmeikapam, Asif Ekbal and Sivaji Bandyopadhyay, "Named Entity Recognition for Manipuri Using Support Vector Machine," 23rd Pacific Asia Conference on Language, Information and Computation, pages 811–818.

[22] Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis and Constantine D. Spyropoulos, "Learning Decision Trees for Named-Entity Recognition and Classification," Institute of Informatics and Telecommunications, NCSR "Demokritos", 15310.

[23] http://en.wikipedia.org/wiki/Natural_language_processing

[24] Kamal Deep and Vishal Goyal, "DEVELOPMENT OF A PUNJABI TO ENGLISH TRANSLITERATION SYSTEM," International Journal of Computer Science and Communication Vol. 2, No. 2, July-December 2011, pp. 521-526.

[25] Yunita Sari, M. Fadzil Hassan, Norshuhani Zamin, "A Hybrid Approach to Semi-Supervised Named Entity Recognition in Health, Safety and Environment Reports,"International Conference on Future Computer and Communication © 2009 IEEE.