# Domain Specific Named Entity Recognition (DSNER) from Web Documents

Pawan kumar
Research Scholar

Raj Kumar Goel
Assistant Professor
Noida Institute of Engineering & Technology, Gr. Noida (UP)

Prem Sagar Sharma
Assistant Professor
J.P. Institute of Engineering & Technology, Meerut (UP)

## ABSTRACT
Named entity recognition is a tool, which use process natural language tasks such as, text categorization, speech translation, and document classification. The Web data promotes the idea, that more and more data can be interconnected. A step towards this goal is to bring more structured annotations to existing documents using common vocabularies or ontology. Semi-structured texts such as scientific, medical, forum and blog posts can hence be semantically annotated. Named Entity (NE) extractors play a key role for extracting structured information by identifying features, also called entities, and by linking them to other web resources by means of typed inferences. Earlier many systems have been developed named entity recognition with substantial success save for the problem of being domain specific and making it difficult to use the different systems across domains. In this paper we introduce specific domain like science, medical and news, named entity recognition. This paper presents a system to recognize the Named Entity from web documents using ontology.

## General terms
Introduction, Ontology, Architecture of Keyword Extractor, Algorithm: Keyword Extraction, Architecture of DSNER, Algorithm: DSNER.

## 1. INTRODUCTION
Task of recognizing and classifying single and multi-word expressions within a document that refer to proper names of any kind (person, organization, location etc.) can be described as Named Entity Recognition (NER). For example, in the text: "Arvind kejriwal is the chief-minister of Delhi on Dec 28, 2013" a NER system should recognize "Arvind kejriwal" as Person Named Entities, "Delhi" as Location Named Entities, "Chief-minister" as Position Named Entity, "Dec 28, 2013" as Date/Time Named Entity. [1] There is a lot of interest in NER as this domain is situated halfway towards understanding of Natural Language. Named Entities respond to the questions like "who?" and "where?". Such kind of answers are much practical for creating any semantic representations of sentences like in the case of Information Extraction systems [2][3] and Human-Machine Dialogue systems or merely for indexing texts like in the case of match calculation between texts [4]. NER can also be very functional in parallel corpora alignment (using Named Entities as anchor points between parallel texts). Ontology represents knowledge as a set of concepts secret a domain, and the relationships between such concepts. It can be used to describe the domain and may be used to reason about the entities within that domain. Considering hypothesis, ontology is a "formal, accurate specification of a shared concept". Shared vocabulary and taxonomy has been released by Ontology which models a domain with the objects definition and/or concepts and their relations and properties. Foundation of the term ontology in philosophy and has been applied in a range of ways. The foundation meaning within computer science is a model for describing the world that consists of a set of types, relationship types and properties. Indeed what is provided around these parts, but they are the requisites of ontology. Generally, there is also an expectation that there be a close resemblance between the features of the model in ontology and the real world.

## 2. RELATED WORK
The goal of the NER focuses at automatically and in large volumes of text which robustly annotating named entities. Being able to adapt to different domains and document genre's without much (or any) tuning, NER systems are required to offer good performance [6]. Previously, various methods have been used by Named Entity Recognition systems. Depending on the particular method, these systems can be classified into three general categories as:

- Systems using hand crafted grammars of recognition like the New York University [6] entrant in MUC-6;
- Systems using "machine learning" techniques like all the systems used in the CoNLL shared task conferences. Some of them did very well on both languages;
- Hybrid systems using both approaches, for example, the very successful system of the Language Technology Group at the University of Edinburgh [7] presented in the MUC-7 conference.

On the other hand Jiang and Zhai [2006] present several strategies for exploit the domain structure in training data to learn a more robust and named entity recognizer that can perform well on any new domain. They cobble together a way to automatically rank features based on how generalizable they are across domains. They then train a classifier with strong emphasis on the most generalizable features [6]. Nadeau et al. [2006] use an un-supervise strategy for domain independence by creating a system that can recognize named-entities in a given document without prior training by using automatically generated gazetteers and later resolving ambiguity [6].

## 3. NAMED ENTITY RECOGNITION (NER) AND ONTOLOGY
One of the first research papers in the NLP field aiming at automatically identifying named entities in texts was proposed by Rau [7] and work relies on heuristics and definition of patterns to recognize company names in texts. Semi-Supervised Learning(SSL) approach and Unsupervised Learning (UL) approach attempt to solve such types of problem by either providing a small initial set of labeled data to train and seed the system [8] or by resolving the extraction problem as a clustering one of these. The training set is defined by the set of heuristics chosen. Moreover the different learning approaches, the Named Entity recognition tools differ in terms of the language they can support here. Even as each

language has its own syntax and semantics that may affect the way the entities can be extracted, Palmer et al. have used statistical methods for finding named entities in newswire articles for English, Chinese, French, Japanese and Spanish [12]. For example, one can try to gather named entities from clustered groups based on the similarity of context where as other unsupervised methods may rely on lexical resources (e.g. WordNet), statistics computed on large annotated corpus and lexical patterns [13]. They found that the complexity of the NER task was different for the six languages other than that a large part of the task could be performed with simple methods. On the other hand, the results were affected by low F-measure and nonexistence of mapping between entities to the types. In this paper, we consider English as a language in order to take out one variable in our evaluation. The NERD framework is conversely independent of the language relying exclusively on the capabilities of the essential named entity extractors. A different approach was introduced when Supervised Learning (SL) techniques were used. The big disruptive change was the use of a large dataset manually labeled. In the SL field, a human being usually trains positive and negative examples in order that the algorithm computes classification patterns. SL techniques exploit Support Hidden Markov Models (HMM) [15], Vector Machines (SVM) [14], Decision Trees, Conditional Random Fields (CRF) [17] and Maximum Entropy Models [16]. The common goal of these approaches is to be familiar with relevant key-phrases and to classify them in fixed taxonomy. The challenges with SL approach the prohibitive cost of creating examples and are the unavailability of such labeled resources. In computer science and information science, ontology represents knowledge as a set of concepts within a specific domain, and the relationships between such concepts. It can be used to reason with reference to the entities within that domain and may be used to describe the domain. In hypothesis, ontology is a "formal, explicit specification of a shared conceptualization". Ontology renders taxonomy and shared vocabulary which models a domain with the definition of objects and/or concepts and their relations and properties. The term ontology has its foundation in philosophy and has been practical in many different ways. The core meaning within computer science is a model for describing the world that consists of a set of types, relationship types and properties. Precisely what is provided around these varies, but they are the nuts and bolts of ontology. There is also generally an

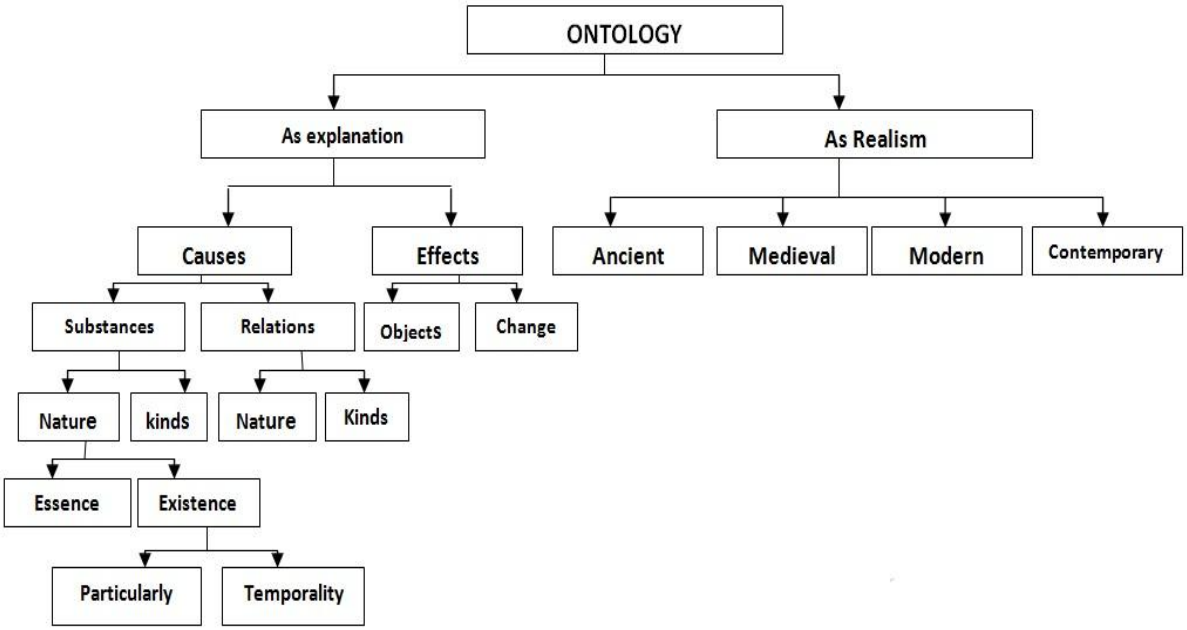


**Fig 1: Representation of ontology**

# 4. PROPOSED WORK

In this paper, proposed a Domain Specific Named Entity Recognizer (DSNER) to extract entities from web documents, in a definite domain. An efficient approach of improving the correctness of NER is done by creating ontology for specific domain and thesaurus.

In this paper enhanced two main systems, keyword extraction as shown in fig-1 and named entity identification as shown fig-2, which both affect on the assessment of system operations.

## 4.1 Components of Keyword Extractor:

Keyword extraction has been categorized into three components as:

a. HTML Cleaning
b. Tokenization
c. Stop word Removal

*a. HTML Cleaning*

Web documents, once downloaded by a web crawler/spider are HTML documents embedding the HTML tags with the content.

For domain exact processing HTML tags has no relevance and hence will be detached from pages before further processing. If the document is a word document/ pdf files etc. then this component will not be use further.

*b. Tokenization*

Tokens are set of characters which are separated by spaces. Documents are the gathering of information in the form ideas and thoughts. Information is represented in the form sentences in a any document. To process these documents, first documents need to tokenize. Following tokenization these tokens are stored in token repository.

c. *Stop Keyword Removal*

Sentences are fashioned using different words called as tokens. Not all of these words are important for making information in a sentences eg- is, am are, will, this, these. Here, such words are called as stop words. These words are detached from token repository. Stop Keyword exclusion takes token from token repository and match it with list in stop keyword dictionary, if match found the token is dropped or else stored in token repository.

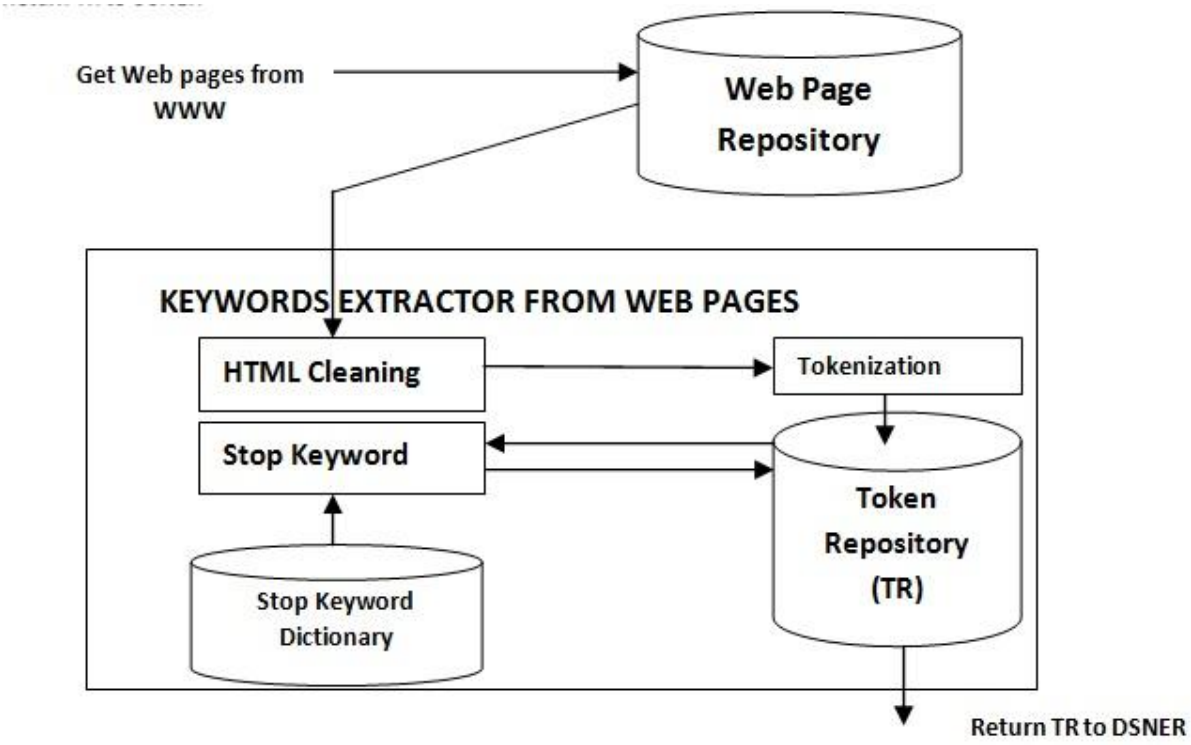


**Fig.2: Architecture of Keyword Extractor**

Algorithm: Keyword Extraction

keywordExtract ()
{
1. [Extract web page one by one from Web Page Repository (WPR) while it is not empty.]
While (WPR [i]! =NULL)
webPage= WPR[i];
2. HTML Cleaning-[Remove all the HTML tags from the web page and convert in to text document]
textDoc = HTMLCleaning (webPage);
3. Tokenization- [Find the Tokens from extracted text document in step-2, and add into Token Repository (TR)]
TR [j++] = Tokenization (textDoc);
4. Stop Keyword – [Remove all the stop keywords from the Token Repository (TR).]
While (TR [i]!=NULL)
Token=TR [i];
If (StopKeword(token)==true)
Remove(token);

5. Return the Token Repository to Domain specific named Entity Recognizer (DSNER).
}

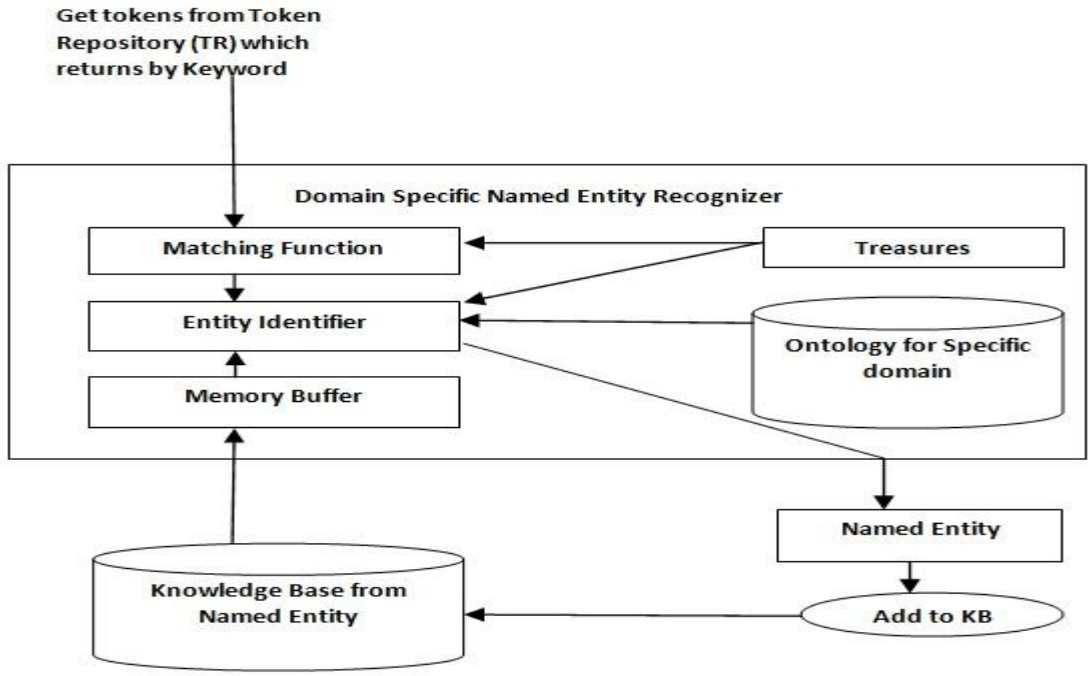


**Fig. 3: Architecture of DSNER**

## 5.1 Components of DSNER

Domain Specific Named Entity Recognizer has been categorized into three components as shown in fig-2 given below:

a. Ontology for specific domain
b. Thesaurus
c. Matching function
d. Entity identifier
e. Knowledge Base

### *a. Ontology for specific domain*

In computer science and information science, ontology formally represents knowledge as a set of concepts within a domain, and the relationships between those concepts.

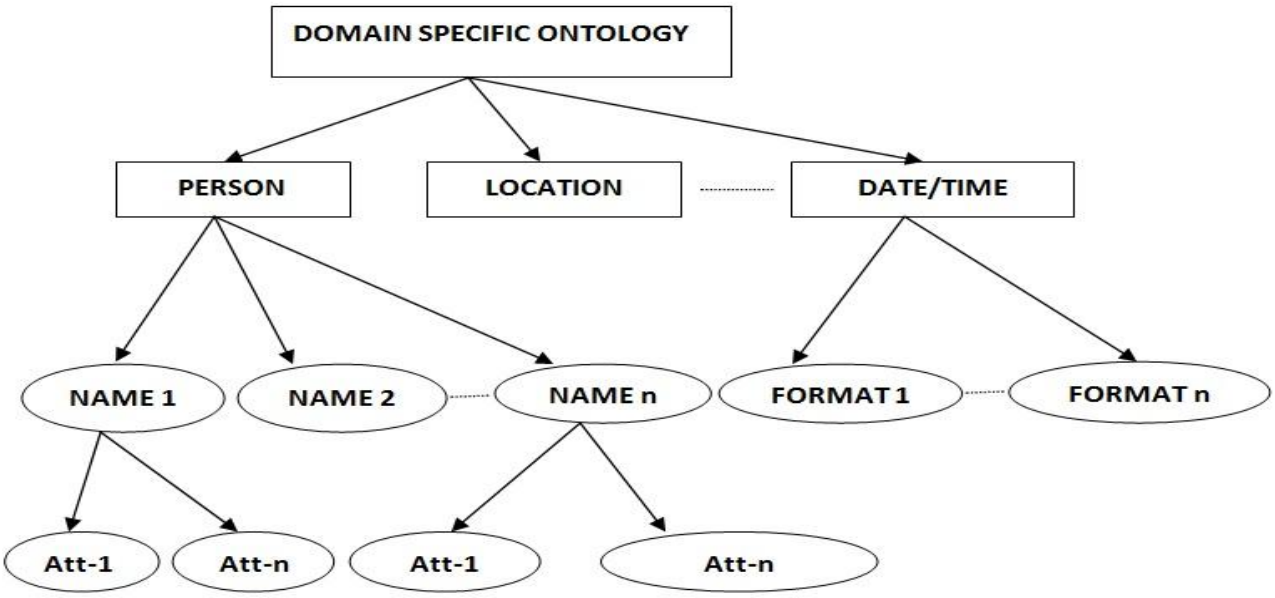


**Fig. 4 Tree ontology for named entity recogazation**

In this paper a tree structure for ontology is chosen. By means of this Ontology we try to convert unstructured web document into structured document.

### *b. Treasures*

To process a document for information different words are encountered. Many of these words are synonyms of any base word. In a document, information is represented using different words which are from time to time synonyms of a

base word. If token is synonyms of a base word subsequently it is convert into Base Token using thesaurus and get ahead of base token into matching function as an argument.

*c. Matching function*

Matching function matches each token in thesaurus and proceeds base token if matched otherwise it come again original token to entity identifier.

*d. Entity Identifier*

Identify the Named Entity consequent to the token using ontology as shown in the fig-4.

*e. Knowledge Base*

One time the Named Entities are identified using Entity Identifier, these are recorded in the knowledge base along by means of document id and token.

Algorithm: Domain Specific Named Entity Recognizer (DSNER)

ds_NER

{

1. Extract token one by one from the Token Repository (TR) which return by Keyword Extractor, while it is not empty.
2. Matching function: Each token match in thesaurus and return base token.

   T= matchFunction (token, thesaurus);
3. Entity Identifier: identify the Named Entity corresponding to the token using ontology.

   if (T != NULL)

   R=entityIdentifer (T, Ontology);

   else

   R= entityIdentifier (token, Ontology);
4. Return the Named Entity with token in the form of <Token, Named Entity>

}

## 5. CONCLUSION & FUTURE SCOPE

The Named Entity Recognition field has been prosperous for more than twenty years. It aims at extracting and classifying mentions of unyielding designators, from text, such as proper names, biological species, and temporal expressions. This work presents a way of extracting named entities from a web document using ontology. The Named Entity Recognition is providing to entity of the web documents and also use in the question answering systems. In the future, we use DSNER in automatic generation question from web documents and then we find the domain of entity. We use the algorithm for keyword extraction from web documents and then we token provide for repository. The Named Entity Recognition extracts the entity from the single-named-entity-queries. Ontology are provide the specific domain entity which entity extract from DSNER. In this paper, we provide entity using ontology from web documents.

## 6. REFERENCES:

[1] Michailidis, Konstantinos Diamantaras & SpirosVasileiadis "Greek Named Entity Recognition using Support Vector Machines", Proceedings of the 7th International Conference on Greek Linguistics

[2] N. Chinchor, "MUC-6 Named Entity Task Definition (version 2.1)", in Proceedings of the 6th Message Understanding Conference (MUC-6), Morgan-Kaufmann, Columbia, Maryland, November 1995.

[3] N. Chinchor, "MUC-7 Named Entity Task Definition (version 3.5)", in Proceedings of the 7th Message Understanding Conference (MUC-7), Fairfax, VA, 19 April - 1 May 1998.

[4] N. Friburger and D. Maurel, "Textual similarity based on Proper Names", in Proceedings of the workshop Mathematical / Formal methods in Information Retrieval (MFIR '2002) at the 25th ACM SIGIR Conference, Tampere, Finland, 2002, pp. 155-167.

[5] N. Friburger and D. Maurel, "Textual similarity based on Proper Names", in Proceedings of the workshop Mathematical / Formal methods in Information Retrieval (MFIR '2002) at the 25th ACM SIGIR Conference, Tampere, Finland, 2002, pp. 155-167.

[6] Kitoogo Fredrick Edward, Venansius Baryamureeba & Guy De Pauw, "Towards Domain Independent Named Entity Recognition" In International Journal Of Computing And Ict Research, Vol. 2, No. 2, December 2008 Pp. 84-95

[7] L.F. Rau. "Extracting company names from text" In 7th IEEE Conference on Artifcial Intelligence Applications, volume i, pages 29{32, 1991.

[8] Satoshi Sekine. NYU: "Description of the Japanese NE system used for MET-2", In7th Message Understanding Conference (MUC-7, 1998.

[9] Satoshi Sekine and Chikashi Nobata. "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy", In 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, 2004

[10] J. Sim and C.C. Wright. "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy", 85(3):257{268, January 2005.

[11] Giuseppe Rizzo and Raphael Troncy, "NERD: Evaluating Named Entity Recognition Tools in the Web of Data"

[12] David Palmer and David Day. "A statistical pro_le of the Named Entity task" In 5th International Conference on Applied Natural Language Processing, pages 190{193, Washington, USA, 1997.

[13] Enrique Alfonseca and Suresh Manandhar. "An Unsupervised Method for General Named Entity Recognition And Automated Concept Discovery" In 1st International Conference on General WordNet, 2002.

[14] Masayuki Asahara and Yuji Matsumoto. "Japanese Named Entity extraction with redundant morphological analysis" In International Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL'03), pages 8{15, Edmonton, Canada, 2003.

[15] Daniel Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: "a high-performance learning name finder" , In 5th International Conference on

Applied Natural Language Processing, pages 194{201, Washington, USA, 1997.

[16] Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. "NYU: Description of the MENE Named Entity System as Used in MUC-7", in 7th Message Understanding Conference (MUC-7), 1998.

[17] Andrew McCallumand, Wei Li. "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons", In 7th International Conference on Natural Language Learning at HLT-NAACL (CONLL'03), pages 188-191, Edmonton, Canada, 2003.