Pertanika J. Sci. & Technol. 26 (4): 1801 - 1822 (2018)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Named Entity Recognition in Hindi Using Hyperspace Analogue to Language and Conditional Random Field

Arti Jain* and Anuja Arora

CSE & IT, Jaypee Institute of Information Technology, Noida, U.P. 201309, India

ABSTRACT

Named Entity Recognition (NER) is defined as identification and classification of Named Entities (NEs) into set of well-defined categories. Many rule-based, machine learning based, and hybrid approaches have been devised to deal with NER, particularly, for the English language. However, in case of Hindi language several perplexing challenges occur that are detailed in this research paper. A new approach is proposed to perform Hindi NE Recognition using semantic properties to handle some of the Hindi language specific NER challenges. And because of increasing demand in Hindi health care applications, Hindi Health Data (HHD) is crawled from four well-known Indian websites: Traditional Knowledge Digital Library; Ministry of Ayush; University of Patanjali; and Linguistic Data Consortium for Indian Languages. Four novel NE types are determined, namely- Person NE, Disease NE, Symptom NE and Consumable NE. For training purpose, HHD data is converted into Hyperspace Analogue to Language (HAL) vectors, thereby, maps each word into a high dimensional space. Conditional Random Field model is applied based on HHD feature engineering, HHD gazetteers and HAL. Blind test data is then mapped into the high dimensional space created during the training phase and outputs the annotated test data. The results obtained are quite significant; and HAL accompanied with CRF approach seems to provide effective outcome for Hindi NE Recognition.

Keywords: Conditional Random Field, Hindi, Hyperspace Analogue to Language, Named Entity Recognition

INTRODUCTION

Named Entity Recognition (NER) (NER) (Nadeau & Sekine, 2007; Ekbal & Bandyopadhyay, 2008; Srivastava et al., 2011; Rodriquez et al., 2012; Marrero et al., 2013; Baldwin et al., 2015; Ekbal et al., 2016; Patil et al., 2016; Baksa et al., 2017) is a non-trivial, automated sequence

ARTICLE INFO

Article history:
Received: 09 January 2018
Accepted: 05 May 2018
Published: 24 October 2018

E-mail addresses:
ajain.jiit@gmail.com (Arti Jain)
'anuja.arora29@gmail.com (Anuja Arora)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

labelling task which comprises identification and classification of Named Entities (NEs). Identification of NE means marking the presence of a word/term/phrase i.e. names (noun/noun phrase) as NE in a given text. And, classification of NE means denoting the role of an identified NE into certain well-defined categories such as Person, Location, Organization, Money, Date and Time. NER is treated as a main sub-task of Information Extraction (IE) (Grishman, 1995; Chinchor & Robinson, 1997) and is successful in vivid application areas such as Question Answering (Khalid et al., 2008), Machine Translation (Aggarwal & Zhai, 2012), Automatic Text Summarization (Gupta & Lehal, 2011), Word Sense Disambiguation (Moro et al., 2014) and so on. In general, there are three main approaches to NER systems, namely- Rule-based approach (Farmakiotou et al., 2000; Chiticariu et al., 2010), Machine learning approach (Jiang et al., 2011; Ekbal et al., 2016), and Hybrid approach (Saha et al., 2008; Rocktäschel et al., 2012). Rule based NER approach comprises of language based hand-crafted rules and other heuristics e.g. set of patterns to classify words for NER system. For this purpose, thorough language knowledge, grammatical expertise and advanced skills related to the language are required to achieve good results. But these rules are non-transferable to other languages and domains. Also, they incur steep maintenance cost especially when new rules are introduced for certain new information or new domain. Machine learning (ML) based NER approach requires huge amount of NE annotated training data to acquire good results. ML further comprises three approaches- Supervised learning (SL), Semi-supervised learning (SSL) and Unsupervised learning (UL). SL involves learning to classify a given set of labelled examples that are made up of the number of features, only when large amount of high quality training data is available e.g. Hiddden Markov Model (Zhou & Su, 2002), Maximum Entropy (Curran & Clark, 2003), Conditional Random Field (Ekbal & Bandyopadhyay, 2009), Support Vector Machine (Saha et al., 2010), and Decision Tree (Szarvas et al., 2006). SSL involves technique such as bootstrapping (Kozareva, 2006) which has a small degree of supervision for starting the learning process. UL involves training with few seed lists and large unannotated corpus where NEs are gathered from cluster groups based on the similarity of context. For example, Collins and Singer (1999) had discussed an unsupervised model for NE classification by the use of unlabelled examples of data. Kim et al. (2002) had proposed an unsupervised NE classification and ensemble technique which used small scale NE dictionary and unlabelled corpus for NEs.

So far, NER system for English language (Grishman, 1995; Rodriquez et al., 2012; Marrero et al., 2013) has already been widely explored. Konkol et al. (2015) had discussed latent semantic based information for NER which considered local context methods and global context methods. Local context uses only a limited context (context window) around the word to infer vector. Most prominent local context methods are: Hyperspace Analogue to Language (HAL), Correlated Occurrence Analogue to Lexical Semantic (COALS),

Random Indexing (RI), Bound Encoding of AggreGate Language Environment (BEAGLE), Purandare and Pedersen (P&P). While global context uses a wider context (whole section or document) around the word to infer vector. Widely used global context methods are: Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA). But these latent semantic based methodologies have been applied for NER in English and some European languages, which can further be explored for Hindi- an Indian language based NER as well. Since NER system for Hindi (Ekbal & Bandyopadhyay, 2009; Saha et al., 2010; Srivastava et al. 2011; Athavale et al., 2016) is still quite challenging.

Formulation of Problem for Hindi NER

Hindi is written in the Devanagari (Gupta et al., 2011) script and is considered as an official language of the Government of India, in addition to English. And outside India, it is an official language in Fiji, and regional language in Mauritius, Trinidad and Tobago, Guyana, and Suriname. Hindi is highly inflectional, morphologically rich and primarily suffixing language. An excellent source for Hindi language processing is the Hindi WordNet (http://www.cfilt.iitb.ac.in/wordnet/webhwn/). From the past few years, Hindi NER task (Cucerzan & Yarowsky, 1999; Li & McCallum, 2003; Ekbal et al., 2008; Saha et al., 2008; Krishnarao et al., 2009; Srivastava et al., 2011; Athavale et al., 2016) is considered as a budding research topic.

In this paper, NER for the Hindi language using machine learning based Hyperspace Analogue to Language (HAL) methodology is proposed. System is trained for Hindi Health Domain (HHD) corpus, large part of which is unlabelled, as input, which is then transformed into feature vectors (representation of words) along with labels (representation of entities). To do so, training data passes through pre-processing (tokenization, fill-in missing values or gaps); HAL steps (word-vector generation, co-occurrence matrix, and similarity measurement); feature engineering (head nouns, word suffix, part-of-speech and n-gram); seed gazetteers and their extensions through Hindi WordNet. Training algorithm then estimates parameters for the Conditional Random Filed (CRF) model using the trained data. After completion of the learning process, the unannotated blind HHD test data is processed and is transformed into feature vectors using HAL model. CRF is then applied and map the test words to HHD NEs. System implementation of the proposed Hindi NER approach is performed using python® (https://www.python.org/downloads/release/python-2711/) and its supportive ML libraries, to develop HAL model and to find NEs for the test data using CRF. The results that are obtained are quite significant, and the proposed approach is novel for the Hindi NER system.

While performing the Hindi language based NER task, several implementation challenges are encountered. The biggest challenge was crawling of Hindi health corpus

from various websites which was quite complex as Hindi language-based sites followed varying Unicode styles and required integration. Lack of certain python libraries to Unicode support; compatibility of Hindi WordNet w.r.to.python; formulation of gazetteer lists due to lack of standardized Hindi health-based gazetteers are few other challenges. The proposed NER approach is applicable on any social media such as Twitter health tweets, Patanjali Ayurvedic site etc., wherever health content is made available in Hindi.

Research Contributions

This research provides three valuable contributions as stated below:

RC1: Explore Hindi health domain based named entities and relevant gazetteers using Hindi lexical resource (Hindi WordNet).

RC2: Propose latent semantics-based Hyperspace Analogue to Language as state-of-art NER technique for Hindi.

RC3: Study the impact of feature engineering, gazetteers and HAL on NER in Hindi and achieve significant results using CRF method.

Organization of Paper

The rest of the paper is organized as follows: Section 2 introduces the chosen training corpus from Hindi health domain (HHD) Indian sites. This section comprises of HHD NEs and HHD gazetteers. Section 3 discusses contemporary challenges in Hindi NER. Section 4 gives detailed architecture of proposed NER system for Hindi. Section 5 illustrates hyperspace analogue to language semantic details using HAL algorithm and HAL example. Section 6 explains machine learning based CRF model. Section 7 describes HHD feature engineering module. Section 8 shows experiments and results of the Hindi NE Recognition system. Finally, Section 9 concludes the paper.

Training Corpus

Due to the growing need of smart health applications (www.onlymyhealth.com; https://pmsma.nhp.gov.in/) in Hindi there is a rapid demand for health related NER system. As far, no standard Hindi health-domain corpus is available, so we have crawled data of 310,530 words from the four well-known Hindi health domain based Indian websites viz. (i) Traditional Knowledge Digital Library (http:// www.tkdl.res.in/), (ii) Ministry of Ayush (http://ayush.gov.in/), (iii) University of Patanjali (https://www.patanjaliayurved.net/), and (iv) Linguistic Data Consortium for Indian Languages (http://www.ldcil.org/). Figure 1 represents sample crawled Hindi Health Domain (HHD) corpus.

In this research work, we have considered four NEs- Person (PER), Disease (DIS), Symptom (SMP) and Consumable (CNS) NEs for HHD corpus. As per the NER research

किसी एक बीमारी से हमारे देश में इतनी मौतें नहीं होतीं जितनी निमोनिया (Pneumonia) से होती हैं।

निमोनिया से बचाव और इसका इलाज बेहद सुगम है लेकिन अकसर लोगों के पास इसके जुड़ी जानकारी नहीं होती।
जानिए निमोनिया के विषय में सभी बातें।

निमोनिया के घरेलू उपचार

- हल्दी, काली मिर्च, मेथी और अदरक जैसे प्रतिदिन उपयोग में आने वाले खाद्य प्रदार्थ फेफड़ों के लिए फायदेमंद होते हैं।
- तिल के बीज भी निमोनिया के उपचार में सहायक होते हैं। 300 मिलीलीटर पानी में 15 ग्राम तिल के बीज, एक चुटकी साधारण नमक, एक चम्मच

- अलसी और एक चम्मच शहद मिलकर प्रतिदिन उपयोग करने से फेफड़ों से कफ बाहर निकलता है।

- ताजा अदरक का रस लेने या अदरक को चूसने से भी निमोनिया में आराम मिलता है।

- थोड़े से गुनगुने पानी के साथ शहद लेना भी लाभदायक रहता है।

- गर्म तारपीन तेल का और कपूर के मिश्रण से छाती पर मालिश करने से निमोनिया से राहत मिलती है।

- रोगी का कमरा स्वच्छ, और गर्म होना चाहिए। कमरे में सूर्य की रौशनी अवश्य आनी चाहिये।

- रोगी के शरीर को गर्म रखें, विशेषकर छाती और पैरों को।

Figure 1. HHD sample crawled data

guidelines that are undertaken by the AU-KBC Research Centre, Chennai, among the four considered NE types, the first two NEs-PER and DIS are direct sub-categories of ENAMEX (ltrc.iiit.ac.in/iasnlp2014/slides/lecture/sobha-ner.ppt), the third NE- SMP is extracted from fine-grained variation of DIS, and the fourth NE- CNS is extracted from Material sub-category of ENAMEX. Further, the current work can be extended with some more NEs that can be made available in the chosen corpus such as Food, Diagnosis, Treatment etc. Presently, all such words are considered as Not-Named Entity (NNE). Figure 2 shows the considered NEs and their integration relationships. Although researchers in the past have identified Person NE in news-wire domain but they have not worked with the rest three NEs because there is no research work being conducted so far on health domain for Hindi NER.

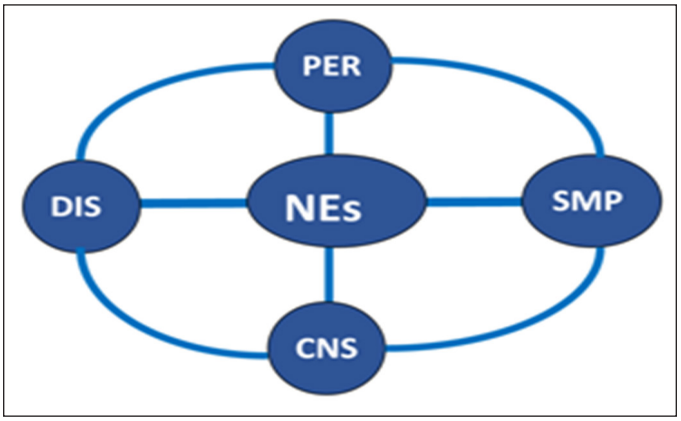

Figure 2. HHD named entities

Description of the selected four NE types for this NER research is given as follows:

Person (PER): PER refers to the person (human being) who may be a single individual or a group. Person entity in HHD corpus contains semantic roles of a person who is directly or indirectly involved in or is affected by certain disease. For Example- "व्यक्ति" (Vyakti/Person), "मरीज़" (Marij/Patient), "महिला" (Mahila/Lady), "आदमी" (Aadmi/Man), "चिकित्सक" (Chikitsak/Doctor), "एक्सपर्ट्स" (Experts) etc.

Disease (DIS): DIS refers to the name of the disease that adversely affects a patient irrespective of the fact that the disease is mild or severe For Example- "दमा" (Dama/Asthama), "चेचक" (Chechak/Chickenpox), "हैजा" (Heja/Cholera), "कालीखांसी" (Kali Khasi/Whooping Cough) etc.

Symptom (SMP): SMP refers to an undesirable physical or mental state of a patient that is regarded as an indicator to some well-known or unknown disease. For Example- "सूजन" (Sujan/Swelling), "मतली" (Matli/Nausea), "पीड़ा" (Peeda/Pain), "इंफेक्शन" (Infection), "तकलीफ" (Takleef/Problem), "थकावट" (Thakawat/Tiredness) etc.

Consumable (CNS): CNS refers to a substance that person intakes through various modes (e.g. oral, inject, inhale, drink, suck, swallow, eat, chew etc.) and is used in pharmacology for diagnosis, prevention, cure or treatment of diseases. For Example- "लहसुन" (Lahasun/Garlic), "दूध" (Dudh/Milk), "एंटीबॉयटिक" (Antibiotic), "ग्लूकोज" (Glucose), "अनाज" (Anaj/Cereals), "राजमा" (Rajama/Beans) etc.

HHD Gazetteers

Gazetteers or gazetteer lists are the entities dictionaries which are important for performing NER effectively (Kazama & Torisawa, 2008; Dey & Prukayastha, 2013; Sahin et al., 2017). They are neither dependent on previously discovered tokens nor on annotations. They only expect a raw text as an input and then find matches based on its contents. In the current work, initial four seed gazetteer lists are chosen manually from HHD corpus. These four lists are having- 107 entries for Person NE e.g. "बच्चा" (Baccha/Child), 141 entries for Disease NE e.g. "गठिया" (Gathiya/Arthiritis), 223 entries for Symptom NE e.g. "दर्द" (Dard/Pain), and 388 entries for Consumable NE (CNS) e.g. "खाना" (Khana/Food) respectively.

Later on, each of these lists is extended through semi-automatic process using Hindi WordNet synset through python®code. As a result, each of these four gazetteers are extended, having 860 entries for PER (e.g. "बच्चा" (Baccha/Child) has extensions as "नवजात_शिशु" (Navajat_shishu/Newborn baby), "नवजातक" (Navajataka/New born), "लड़का"

(Ladka/Boy), "बालक" (Balak/Boy), "छोकड़ा" (Chhokadaa/Man-child), "छोरा" (Chhora/Lad), "छोकरा" (Chhokara/Chap), "लौंडा" (Launda/Bugger), "वत्स" (Vats/Child), "नन्हा-मुन्ना" (Nanha-munna/Child), "नन्हा_मुन्ना" (Nanha_munna/Child), "पुत्र" (Putr/Son), "बेटा" (Beta/Son), "सुत" (Sut/Son), "शिशु" (Shishu/Baby) etc.); 597 entries for DIS (e.g. "गठिया" (Gathiya/Arthritis) has extensions as "संधिवात" (Sandhivata), "संधिशोथ" (Sandhishoth), "सन्धिवात" (Sandivata), "सन्धिशोथ" (Sandhishoth), "संधि_शूल" (Sandhi_Shula), "डमरुआ" (Damruaa), "डबरुआ" (Dabruaa), "पवन-व्याधि" (Pawan-vyadhi), "आर्थ्राइटिस" (Arthritis), "आर्थराइटिस" (Arthritis) etc.); 2655 entries for SMP (e.g. "दर्द" (Dard/Pain) has extensions as "तकलीफ" (Takleef), "दरद" (Darad), "पीड़ा" (Peeda), "तक़लीफ़" (Takleef), "पीर" (Pir), "हूक" (Huuk), "उपताप" (Uptap), "उत्ताप" (Utap), "पीरा" (Pira), "वेदना" (Vedana), "बेदना" (Bedana), "क्लेश" (klesh), "व्यथा" (Vyatha), "अनुसाल" (Anusal) etc.); and 3828 entries for CNS (e.g. "खाना" (Khana/Food) has extensions as "खाद्य_वस्तु" (Khady_vastu), "खाद्य_पदार्थ" (Khady_padarth), "आहार" (Ahar), "खाद्य" (Khady), "भोज्य_पदार्थ" (Bhojy_padarth), "खाद्य_सामग्री" (Khady_samagri), "अन्न" (Ann), "आहर" (Aahr), "फूड" (Food), "भोजन" (Bhojan), "रसोई" (Rasoi), "रोटी" (Roti), "डाइट" (Diet) etc.) respectively.

Contemporary Challenges in HHD NER

NER for humans appear to be straightforward as most of the NEs are the proper names. But for a machine to learn and understand NER is comparatively hard, especially for Hindi. A few researchers have identified challenges in Hindi NER (Ekbal et al., 2016; Jain et al., 2014; Saha et al., 2012; Srivastava et al., 2011). Some new and previously mentioned Hindi NER challenges are listed below:

Rare occurrence of certain NEs in HHD corpus: e.g. "कुटकी चिरौता" (Kutki Chirauta) which is a NE under CNS and has rare occurrence in HHD corpus.

Multiple ways of mentioning the same NE:

Variation in PER semantic information: e.g. "रोगी" (Rogi), "पेशंट" (Patient), "पेशेंट" (Patient), "मरीज" (Marij), "मरीज़" (Marij) all refer to PER (Patient).

Variation in DIS semantic information: e.g. "डायबीटीज़" (Diabetes), "डायबीटीज" (Diabetes), "डायबिटीज़" (Diabetes), "डायबिटीज" (Diabetes), "मधुमेह" (Madhumeh), "मधुप्रमेह" (Madhuprameh), "इक्षुप्रमेह" (Ikshuprameh), "मूत्रकृच्छ" (Mutrakarachchh) all refer to the DIS (Diabetes).

Variation in SMP semantic information: e.g. "झुनझुनी" (Jhunjhuni), "झनझनाहट"

(Jhnajhannahat), "झुनझुनाहट" (Jhunjhunahat), "सुरसुरी" (Sursuri), "सनसनाहट" (Sansanahat), "सनसनी" (Sansani), "सनसन" (Sansan), "सन-सन" (San-san) all refer to SMP (Tingle).

Variation in CNS semantic information: e.g. "कॉफ़ी" (Coffee), "कॉफी" (Coffee), "काफी" (Coffee), "काफ़ी" (Coffee), "कॉफ़ीपावडर" (Coffee Powder), "कॉफीपावडर" (Coffee Powder), "काफीपावडर" (Coffee Powder), "काफ़ीपावडर" (Coffee Powder) all refer to CNS (Coffee).

Disease vs. Symptom: e.g. "बदहजमी" (Badhazmi/Indigestion), "ज़ुकाम" (Jukam/Colds) refer to DIS NE or SMP NE.

Lack of Capitalization: English language uses capitalization as a discriminating feature for classifying words as NEs. On the other hand, Hindi does not have the concept of capitalization at all. For example, Tuberculosis (T.B.) is a Disease in English and is represented as "टी. बी." (T.B.) in Hindi. Similarly, "एड्स" (AIDS), "विटामिनई" (Vitamin E) etc.

Lack of well-defined Gazetteers: Well-defined NE gazetteers are not freely available for Hindi.

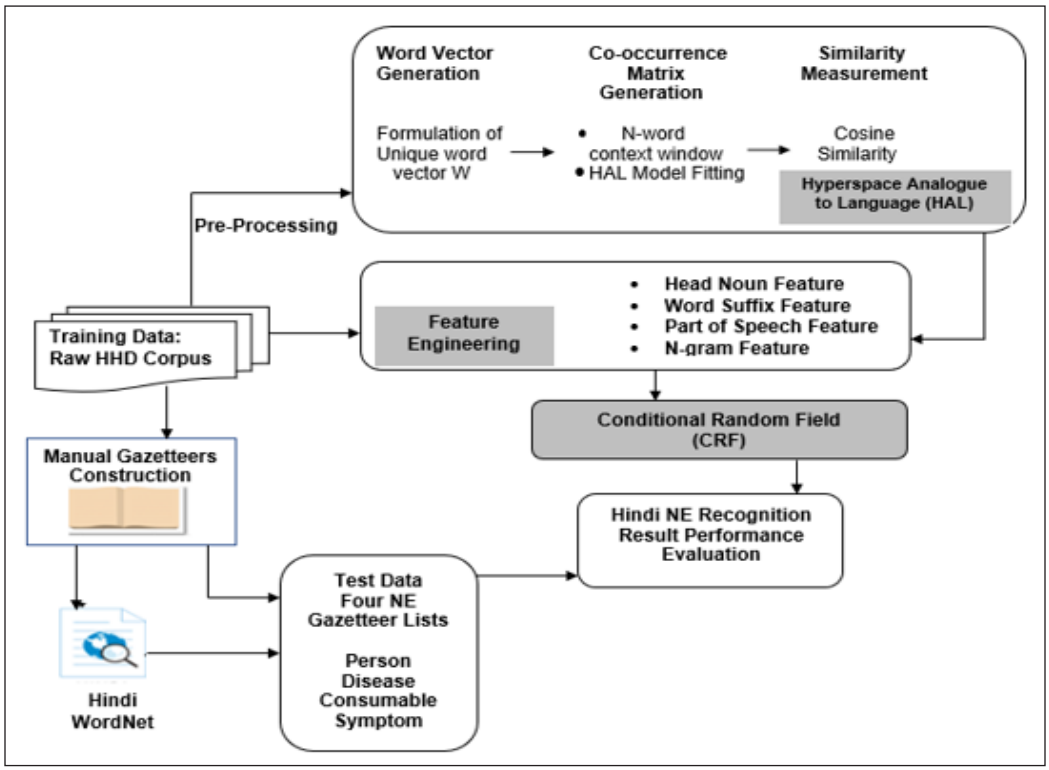

Figure 3. NER system architecture

Proposed NER System for Hindi

Architecture of HHD based NER system is depicted in Figure 3. This system works into training and test phases as follows. The training phase takes the annotated training HHD corpus and is then transformed into feature vectors (representation of words) along with labels (representation of entities). For this purpose, training data passes through HAL steps (word vector generation, co-occurrence matrix, and similarity measurement); feature engineering (head nouns, word suffix, part-of-speech and n-gram); manual gazetteer construction and their extension through Hindi WordNet. The training algorithm then estimates parameters for the CRF model using this trained dataset. Now the unannotated test data is supplied to the NER system and is transformed into feature vectors, CRF is applied onto this data and results into output annotation for the test data.

We have used nltk-3.2.4 (https://pypi.python.org/pypi/nltk), hal-0.0.3 (https://pypi.python.org/pypi/HAL), scikit-learn 0.15.2 (https://pypi.python.org/pypi/scikit-learn/0.15.2) as set of python modules for the Hindi NER task.

Hyperspace Analogue to Language

Hyperspace Analogue to Language (HAL) (Tayal et al., 2015) is also known as semantic memory and was developed by Kevin Lund and Curt Burgess, University of California, Riverside, California in 1996. HAL basic premise (Lund & Burgess, 1996; Lund et al., 1996; Burgess & Lund, 1997) relies on the fact that the words with similar meanings repeatedly occur closely (also known as co-occurrence). Another researcher (Firth, 1957) stated that a word is characterized by the company it keeps i.e. meaning of a word can be interpreted by its surrounding contexts, around which that word often appears. In this research, HAL is chosen as computational model that exploits statistics for the contexts of HHD corpus words.

HAL determines the similarities between the words while collecting the statistics about the word co-occurrences, using two vital assumptions-

Left and right context of a word holds different information, and so it is important to keep this statistic as separate entity;

Distance between the words within a sentence is important, and so more distant words are less informative while lesser distant words are more informative.

Such statistics is useful to generate high-dimensional vectors, where each vector represents meaning for one word; and the words that are represented as vectors formulate the vector space model. Then compare the words and their meanings using similarity/distance among vectors. For this purpose, HAL uses local context, also called as limited context or context window around a word to infer its vector. Such a context window contains only a few words before and after the processed word. Thus, HAL is treated as a semantic space model which discovers different kind of relations between words. Consider

an example word "रोग"(Rog/Disease) then HAL aids in finding the local context for this word as- "व्याधि" (Vyadhi/Illness) and "बीमारी"(Bimari/Disease) words that are observed to be the most similar words corresponding to the given example word.

HAL Algorithm

This section discusses about HAL algorithm and its execution through an illustrative example.

Algorithm: HAL Algorithm //python nltk, sklearn implementation

Input Hindi HHD Corpus
Output Latent semantics class for HHD corpus into four HHD NEs
Declare s: number of HHD corpus sentences
w: HHD word
N: number of unique HHD words
W: word vector of unique words, $W = (w_1, w_2, \ldots w_N)$
M: HHD co-occurrence matrix, generated from |W|
i, j: incremental count variable, initialize as 1
k: array index of word vectors
R_v: preceding context of word vector (row wise)
C_v: followed/succeeding context of word vector (column wise)

Begin:
// Process 1: Generate Word Vector (W)
1. For each pre-processed sentence E
2. W = set(E) // formulate unique word vector
3. EndLoop

// Process 2: Generate HHD Co-Occurance Matrix(M)
4. HAL_model = CountVectorizer(ngram_range = ([1: n])) //generate n-word context window
5. W_Trans = HAL_model.fit_transform(W)
//W_Trans: transformed matrix of W
6. M = ((W_Trans.T) *(W_Trans)) //T: transpose
// Process 3: Similarity Measurement
7. vectors = get_vectors (M, W)
8. LOOP (k<=W, i<=W, j<=W)
9. $R_v[k]$ = vectors (W_{ij}) //construct R_v based on preceding contexts
10. EndLOOP
11. LOOP (k<=W, j<=W, i<=W)
12. $C_v[k]$ = vectors (W_{ij}) //construct C_v based on followed contexts
13. EndLOOP
14. LOOP(k<=W)
15. D[k] = $R_v[k]$ + $C_v[k]$ //construct vectors to compute cosine similarity
16. EndLOOP
17. Cm(D[k_i]D[k_j]) = co_sim(D[k_i], D[k_j])
=dot(D[k_i]D[k_j])/(sqrt(dot(D[k_i]D[k_j])) *sqrt(dot(D[k_j]D[k_j])))
// cosine similarity between two-word vectors
End

Figure 4. HAL Algorithm for Hindi NER

HAL Illustrative Example

Input Hindi Sentences:

"घुटनों के दर्द की सबसे बड़ी वजह है ओवरवेट, जाहिर है आपके भार को सहने में घुटनों को तकलीफ होगी"

(Ghutno ke dard ki sabse badi vajah he overweight, jahir he aapke bhar ko sahane me ghutno ko takleef hogi/ The most common cause of pain in the knees is overweight, obviously will hurt the knee to bear your burden).

The word co-occurrence frequency matrix (M)- a square matrix is generated based upon the number of processed input words. Here, M_{18*18} is constructed as is presented in Table 1. Within the co-occurrence matrix a context window size (n = 5) is chosen to find out the semantic relationships among the words of the matrix. Based on experimental observations, it is seen that if n > 5 then M is highly sparse, words may appear semantically too far from each other. If n < 5 then M is highly dense, moreover, difficult to compute similarities among words, and they may appear too close to each other. Thus, experimental evaluation interprets that the best context size is 5 (n = 5) for words (immediate to distant words, range highest to lowest as 5 to 1) with respect to the HHD corpus.

For an input Hindi word "दर्द" (Dard/Pain), compute preceding context word vector (R_v), succeeding context word vector (C_v) and D[k] as a concatenation of word vector based on R_v and C_v as follows:

$R_v[1]$ = [4 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

$C_v[1]$ = [0 0 0 5 4 3 2 1 0 0 0 0 0 0 0 0 0 0]

D[1] = [4 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5 4 3 2 1 0 0 0 0 0 0 0 0 0 0]

Consider another word "तकलीफ" (Takleef/Problem) Then

$R_v[2]$ = [4 0 0 0 0 0 0 0 0 0 0 0 0 6 2 3 0 0]

$C_v[2]$ = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5]

D[2] = [4 0 0 0 0 0 0 0 0 0 0 0 0 6 2 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5]

Further, Cosine Similarity (Cm) has been computed between two input words using equation (1)

$$Cm(D[1], D[2]) = \frac{D[1].D[2]}{|D[1]| * |D[2]|} \quad (1)$$

So, the following computations are performed as mentioned below:

D[1]. D[2] = 16

|D[1]| = sqrt((4)^2 + (5)^2 + (5)^2 + (4)^2 + (3)^2 + (2)^2 + (1)^2) = 9.798

Table 1
HAL co-occurrence matrix

	घुटनों	के	दर्द	की	सबसे	बड़ी	वजह	है	ओवरवेट	,	जाहिर	आपके	भार	को	सहने	में	तकलीफ	होगी
घुटनों	0	0	0	0	0	0	0	0	0	0	0	1	2	3	4	5	0	0
के	5	0	0	0	0 0	0	0	0	0	0	0	0	0	0	0	0	0	0
दर्द	4	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
की	3	4	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
सबसे	2	3	4	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
बड़ी	1	2	3	4	5	0	0	0	0	0	0	0	0	0	0	0	0	0
वजह	0	1	2	3	4	5	0	0	0	0	0	0	0	0	0	0	0	0
है	0	0	1	2	3	4	5+1	2	3	4	5	0	0	0	0	0	0	0
ओवरवेट	0	0	0	1	2	3	4	5	0	0	0	0	0	0	0	0	0	0
,	0	0	0	0	1	2	3	4	5	0	0	0	0	0	0	0	0	0
जाहिर	2	0	0	0	0	1	2	3	4	5	0	0	0	0	0	0	0	0
आपके	0	0	0	0	0	0	0	5+1	2	3	4	0	0	0	0	0	0	0
भार	0	0	0	0	0	0	0	4	1	2	3	5	0	0	0	0	0	0
को	5	0	0	0	0	0	0	3	0	1	2	4	5+1	2	3	4	0	0
सहने	0	0	0	0	0	0	0	2	0	0	1	3	4	5	0	0	0	0
में	0	0	0	0	0	0	0	1	0	0	0	2	3	4	5	0	0	0
तकलीफ	4	0	0	0	0	0	0	0	0	0	0	0	0	5+1	2	3	0	0
होगी	3	0	0	0	0	0	0	0	0	0	0	0	0	4	1	2	5	0

|D[2]| = sqrt((4)^2 + (6)^2 + (2)^2 + (3)^2 + (5)^2) = 9.487

Therefore, Cosine Similarity among words "दर्द" (Dard/Pain) and "तकलीफ" (Takleef/Problem) is Cm(D[1], D[2])=0.172.

In addition, cosine similarity among words "दर्द" (Dard/Pain) and "ओवरवेट" (Overweight) can be computed out to be 0.029, while cosine similarity among words "तकलीफ" (Takleef/Problem) and "ओवरवेट" (Overweight) can be computed out to be 0.010. These observations clearly indicate that the two words "दर्द" (Dard/Pain) and "तकलीफ" (Takleef/Problem) are semantically closer to each other, while "दर्द" (Dard/Pain) and "ओवरवेट" (Overweight); "तकलीफ" (Takleef/Problem) and "ओवरवेट" (Overweight) are not semantically close to each other.

Conditional Random Field Framework

Conditional Random Field (CRF) (Li & McCallum, 2003) is a probabilistic based discriminative, undirected graphical model that is highly useable for sequential labelling tasks such as part-of-speech tagging (PVS & Karthik, 2007), table extraction (Pinto et al., 2003), named entity recognition (Ekbal & Bandyopadhyay, 2009), noun phrase segmentation (Sha & Pereira, 2003). CRF has the capability to easily add-on large number of arbitrary, non-independent features in conjunctions to the base features. CRF calculates the conditional probability values on the designated output nodes, given values as are assigned to other designated input nodes.

CRF defines the conditional probability of state sequence $s = <s_1, s_2, s_3...s_T>$, given an observation sequence $o = <o_1, o_2, o_3...o_T>$ as in equation (2):

$$P(s \mid o) = \frac{1}{z_o} exp \sum_{t=1}^{T} \sum_{k=1}^{M} \lambda_k f_k(s_{t-1}, s_t, o, t) \quad (2)$$

Here,

T : number of tokens in a sequence

M : number of features

$f_k(s_{t-1}, s_t, o, t)$: feature function, weight λ_k is learnt via training

z_o : normalization factor over all state sequences

The values of the feature functions may range between $-\infty$, $+\infty$ but typically they are binary. Under binary, $f_k(s_{t-1}, s_t, o, t)$ has value of 0 for most cases, and is only set to be 1 when s_t , s_t are certain states and the observation has certain properties.

Also, to make all conditional probabilities sum up to 1, z_o is defined as in equation (3):

$$z_o = \sum_s \exp(\sum_{t=1}^{T}\sum_{k=1}^{M} \lambda_k f_k(s_{t-1}, s_t, o, t)) \quad (3)$$

In order to train CRF, objective function to be maximized is the penalized log-likelihood of the state sequences, given the observation sequences as in equation (4):

$$L = \sum_{i=1}^{N} \log(P(s^{(i)} \mid o^{(i)}) - \sum_{k=1}^{M} \frac{\lambda_k^2}{2\sigma^2} \quad (4)$$

where,

$\{< o^{(i)}, s^{(i)} >\}$: labelled training data with observed sequence as tokens and state sequence as corresponding labels

$\sum_{k=1}^{M} \frac{\lambda_k^2}{2\sigma^2}$: sum which corresponds to zero-mean

σ^2 : variance, Gaussian prior for parameters optimization

We have used *sklearn-crfsuite 0.3*.6 (https://pypi.python.org/pypi/sklearn-crfsuite) as an open source implementation of CRF for segmenting or labelling sequential HHD corpus.

HHD Feature Engineering

This section details about varied HHD corpus-based features that are used for the experiments.

Head noun feature: Head noun feature (F_{hn}) is usually defined as a major noun or noun phrase of an NE which describes its function or property. It serves as unigram, bigram and trigram head nouns.

Word suffix feature: Word suffix feature (F_{ws}) represents suffix of the current and/or surrounding words. Currently, length of 2 to 4 characters is used as feature. Table 2 shows sample suffixes along with examples from HHD corpus.

Part-of-speech feature: Part-of-speech (POS) feature (F_{ps}) represents the POS information for the previous words and the current word of HHD corpus using POS tagger (https://bitbucket.org/sivareddyg/hindi-part-of-speech-tagger). Several coarse-grained POS tags, such as NomPSP which represents nominal followed by a post-position marker is considered.

N-gram feature: N-gram feature (F_{ng}) extracts n-tuple of HHD corpus words. In this research, only bi-grams and tri-grams are considered while other higher order n-grams are

restricted because of the limitation in the size of the HHD corpus. Table 3 shows some examples of bi-gram and tri-gram features, for all the four HHD corpus-based NE types.

Table 2
Word suffixes and examples

Suffix	HHD Examples
~दर्द	"सिरदर्द", "पेटदर्द", "गलादर्द", "कमरदर्द"
~हट	"अकुलाहट", "मिचलाहट", "झनझनाहट", "सरसराहट", "खरखराहट"
~पन	"दुबलापन", "गंजापन", "चिपचिपापन", "चिड़चिड़ापन", "भारीपन"
~इटिस	"टेन्टीनाइटिस", "बरसाइटिस", "अर्थराइटिस", "आस्टियोअर्थराइटिस"
~त्सक	"चिकित्सक", "मनोचिकित्सक", "दंतचिकित्सक"
~पान	"धूम्रपान", "खानपान"

Table 3
Bi-gram and Tri-gram examples

NE	Bi-gram Examples	Tri-gram Examples
PER	"दंत चिकित्सक" (Dant Chikitsak/Dentist) "वाहन चालक" (Vahan Chalak/Driver)	"हड्डी रोग विशेषज्ञ" (Haddi Rog Visheshagya/ Orthopedic)
DIS	"दमा रोग" (Dama Rog/Asthma) "प्रोस्टेट कैंसर" (Prostate Cancer)	"घुटनों का दर्द" (Ghutno Ka Dard/Knee Pain) "पित्ताशय की पथरी" (Pittashay Ki Pathri/ Gallbladder Stone)
SMP	"खट्टी डकारें" (Khatti Dakare/Belch) "मामूली चोट" (Mamuuli Chott/Minor Injury)	"खट्टी-खट्टी डकारें" (Khatti Khatti Dakare/Belch) "घुटने में सूजन" (Ghutane Me Sujan/Knee Swelling)
CNS	"काले चने" (Kale Chane/Chickpea) "काली मिर्च" (Kali Mirch/Pepper)	"सूखे हरे मटर" (Sukhe Hare Matar/Dry Green Peas) "मूंग की दाल" (Moong Ki Daal/Yellow Lentil)

RESULTS AND DISCUSSION

The proposed system is evaluated using 25K, 50K and 75K HHD blind test corpus; and precision, recall and F-measure metrics are computed. It is observed that as testing goes beyond 75K then there is a stagnation in the performance of the evaluation metric parameters. It so happens because of the occurrence of the overfitting criteria. Overfitting means failing to fit an additional data or fail to reliably predict future observation which arises as the proposed methodology starts memorizing, rather than learning from the HHD

corpus. Table 4 shows the F-measure values for different feature sets in the proposed Hindi NER system. And, to compute F-measure, following categories are considered:

True Positive (TP): system finds NE and is also marked by human annotator.

False Positive (FP): system finds NE but is not marked by human annotator.

True Negative (TN): system does not find NE and is not marked by human annotator.

False Negative (FN): system does not find NE but is marked by human annotator

Hence, precision is the fraction of the correct NE annotations, and is defined as in equation (5):

$$\text{Precision (P)} = \frac{TP}{TP + FP} \tag{5}$$

Recall is the fraction of the NEs that are successfully annotated, and is defined as in equation (6):

$$\text{Recall (R)} = \frac{TP}{TP + FN} \tag{6}$$

F-measure is the weighted harmonic mean of precision and recall, and is defined as in equation (7):

$$\text{F-measure (F)} = \frac{2.P.R}{P + R} \tag{7}$$

While experimenting with various features under 25K, 50K and 75K blind test corpus, it is observed that F_{hn} feature provides lowest F–values for all four NE types as is seen in F1. As features are added such as F_{ws}, F_{ps} to F_{hn} F-values also increases as is seen in F2 and F3 respectively but further adding F_{ng} feature leads to decrease in F-value for SMP NE type on 50K and 75K both, while rest other NEs F-values keeps increasing as seen in F4. When gazetteer lists (F_{gs}) are added to baseline features (F_{hn}, F_{ws}, F_{ps}, F_{ng}) then F-values increase drastically for all four NE types as is seen in F5. HAL is applied for semantic similarity then F_{hl} alone has slight increase in F-values for 25K PER NE but decrease in PER NE for 50K and 75K both as seen in F6. When F_{gs} is accompanied to F_{hl} then again there is a high increase in F-values for all NE types among 25K, 50K and 75K as is seen in F7. F8 and F9 show different combinations of baseline features along with F_{hl} with variations in F-values for different NE types. DIS NE decrease from 84.96 to 84.34 on 50K, CNS NE somewhat increase from 84.04 to 84.57 on 75K, while rest NEs increase in high amount on varied corpus sizes. Finally, F_{hl} along with baseline and F_{gs} give best result for all NE types on 25K, 50K and 75K respectively. F10 shows NEs best F-values, achievable on 75K test as- 90.69%, 89.09%, 87.84% and 88.93% for PER, DIS, SMP and CNS NE types respectively.

Table 4
F-measure for different features in Hindi NER

Feature ID	Feature(s)	F-Measure (%)											
		25K				50K				75K			
		PER	DIS	SMP	CNS	PER	DIS	SMP	CNS	PER	DIS	SMP	CNS
F1	F_{hn}	60.13	61.34	50.04	50.26	62.14	61.76	51.32	51.19	63.12	62.40	52.40	51.37
F2	F_{hn}, F_{ws}	63.12	63.06	54.33	55.11	64.29	64.15	54.78	55.29	65.22	64.56	55.35	55.83
F3	F_{hn}, F_{ws}, F_{ps}	66.55	68.13	62.43	60.15	66.76	69.01	63.66	61.75	67.45	69.33	64.41	62.57
F4	F_{hn}, F_{ws}, F_{ps}, F_{ng}	74.07	76.23	62.55	66.58	74.67	76.92	62.98	67.03	75.76	77.21	63.52	67.35
F5	F_{hn}, F_{ws}, F_{ps}, F_{ng}, F_{gs}	80.18	80.01	72.56	72.36	81.67	80.10	73.10	73.29	82.86	80.13	73.67	74.62
F6	F_{hl}	80.21	81.49	78.04	76.86	80.35	81.56	78.37	77.12	80.79	81.97	78.64	77.78
F7	F_{gs}, F_{hl}	81.39	82.58	80.32	80.01	82.08	83.42	80.64	80.11	82.94	83.84	80.73	80.15
F8	F_{ws}, F_{ps}, F_{ng}, F_{hl}	81.45	83.86	82.45	80.01	81.55	84.96	83.84	80.10	81.52	85.62	84.04	80.01
F9	F_{hn}, F_{ws}, F_{ps}, F_{ng}, F_{hl}	85.45	84.12	84.57	82.16	86.64	84.34	84.57	82.16	87.47	84.28	84.57	82.16
F10	F_{hn}, F_{ws}, F_{ps}, F_{ng}, F_{gs}, F_{hl}	89.08	88.89	86.68	87.13	89.66	88.23	86.38	87.59	90.69	89.09	87.84	88.93

F_{hn}: head noun feature, F_{ws}: word suffix feature, F_{ps}: POS feature, F_{ng}: n-gram feature, F_{gs}: gazetteer lists, F_{hl}: HAL semantic similarity

It is then observed that the overall F-score of different NE types- Person NE, Disease NE, Symptom NE, and Consumable NE for the proposed Hindi NER technique are- 76.98%, 77.42%, 71.57%, and 71.96% respectively which are quite significant as compared with the Maximum Entropy (ME) model (Ahmed & Sathyaraj, 2015: Chieu & Ng, 2002; Curran & Clark, 2003; Hasanuzzaman et al., 2009; Saha et al., 2009, 2008) on the considered Hindi health domain corpus as is seen in Table 5.

Table 5
Comparison of proposed Hindi NER technique w.r.t. Maximum Entropy Model

NE TYPE	Hindi NER Techniques (F-Measure)	
	Proposed Technique	**Maximum Entropy**
PER	76.98%	76.89%
DIS	77.42%	65.34%
SMP	71.57%	53.26%
CNS	71.96%	55.99%

CONCLUSION AND FUTURE SCOPE

In this research work, NER technique for Hindi language using Hyperspace Analogue to Language (HAL) is proposed. HAL uses the semantic based context knowledge which is vital to determine NEs. Such semantics is exploited by the word similarity based on the semantic spaces to cluster words. Four NE types are determined on Hindi health domain (HHD) corpus viz. Person NE, Disease NE, Symptom NE and Consumable NE. Training data passes through HAL steps (word vector generation, co-occurrence matrix, and similarity measurement); feature engineering (head nouns, word suffix, POS and n-gram); manual gazetteer construction and their extension through Hindi WordNet. The training algorithm then estimates parameters for the Conditional Random Field using the trained dataset. Unannotated test data is supplied to the NER system and is transformed into feature vectors for output annotations of the test data. We have used nltk-3.2.4, hal-0.0.3, scikit-learn 0.15.2, sklearn-crfsuite 0.3.6 as set of python modules for the NER task. NEs best F-values, 75K test, achieves 90.69% for Person NE; 89.09% for Disease NE; 87.84% for Symptom NE; 88.93% for Consumable NE respectively. It is observed that the overall F-measure of different NEs on the proposed Hindi NER technique are quite significant as compared to Maximum Entropy model. In future, we intend to focus on the following:

- HHD corpus can be extended to larger extent so that overfitting issue can be resolved to better extent;
- Recognition of some more NE types such as Food, Diagnosis, Treatment;

- Other local semantic techniques such as Correlated Occurrence Analogue to Lexical Semantic (COALS), Random Indexing (RI) can be explored;
- Global context and semantics through Latent Dirichlet Allocation (LDA) can be taken into consideration to enrich word clusters that will lead to better NE accuracy.

REFERENCES

Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining Text Data*. London: Springer Science & Business Media.

Ahmed, I., & Sathyaraj, R. (2015). Named entity recognition by using maximum entropy. *International Journal of Database Theory and Application, 8*(2), 43-50.

Athavale, V., Bharadwaj, S., Pamecha, M., Prabhu, A., & Shrivastava, M. (2016, November). *Towards deep learning in Hindi NER: An approach to tackle the labelled data scarcity*. Retrieved 16 August, 2017, from https://aclweb.org/anthology/W/W16/W16-6320.pdf

Baksa, K., Golović, D., Glavaš, G., & Šnajder, J. (2017). Tagging named entities in Croatian tweets. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, *4*(1), 20-41.

Baldwin, T., Kim, Y. B., De Marneffe, M. C., Ritter, A., Han, B., & Xu, W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text* (pp. 126-135). Beijing, China.

Burgess, C., & Lund, K. (1997). Representing abstract words and emotional connotation in a high-dimensional memory space. In *Proceedings of the Cognitive Science Society* (pp. 61-66). Hillsdale, New Jersey.

Chinchor, N., & Robinson, P. (1997, September). MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding* (Vol. 29, pp. 1-21). Fairfax, Virginia.

Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., & Vaithyanathan, S. (2010, October). Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1002-1012). Cambridge, Massachusetts.

Chieu, H. L., & Ng, H. T. (2002, August). Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics* (pp. 1-7). Taipei, Taiwan.

Collins, M., & Singer, Y. (1999, June). Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, *EMNLP* (pp. 100-110). PG County, USA.

Cucerzan, S., & Yarowsky, D. (1999, June). Language independent named entity recognition combining morphological and contextual evidence. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (pp. 90-99). PG County, USA.

Curran, J. R., & Clark, S. (2003, May). Language independent NER using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* (Vol. 4, pp. 164-167). Edmonton, Canada.

Dey, A., & Prukayastha, B. S. (2013). Named entity recognition using gazetteer method and n-gram technique for an inflectional language: a hybrid approach. *International Journal of Computer Applications*, *84*(9), 31-35.

Ekbal, A., & Bandyopadhyay, S. (2008). A web-based Bengali news corpus for named entity recognition. *Language Resources and Evaluation, 42*(2), 173-182.

Ekbal, A., & Bandyopadhyay, S. (2009). A conditional random field approach for named entity recognition in Bengali and Hindi. *Linguistic Issues in Language Technology*, *2*(1), 1-44.

Ekbal, A., Haque, R., Das, A., Poka, V., & Bandyopadhyay, S. (2008, January). Language Independent Named Entity Recognition in Indian Languages. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages* (pp. 33-40). Hyderabad, India.

Ekbal, A., Saha, S., & Sikdar, U. K. (2016). On active annotation for named entity recognition. *International Journal of Machine Learning and Cybernetics*, *7*(4), 623-640.

Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000, September). Rule-based named entity recognition for Greek financial texts. In *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)* (pp. 75-78). Patras, Greece.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis* (pp. 1–32). Blackwell, Oxford.

Grishman, R. (1995, November). The NYU System for MUC-6 or Where's the Syntax. In *Proceedings of the 6th Conference on Message Understanding* (pp. 167-175). Columbia, Maryland, USA.

Gupta, J. P., Tayal, D. K., & Gupta, A. (2011). A TENGRAM method-based part-of-speech tagging of multi-category words in Hindi language. *Expert Systems with Applications*, *38*(12), 15084-15093.

Gupta, V., & Lehal, G. S. (2011). Named entity recognition for Punjabi language text summarization. *International Journal of Computer Applications*, *33*(3), 28-32.

Hasanuzzaman, M., Ekbal, A., & Bandyopadhyay, S. (2009). Maximum entropy approach for named entity recognition in Bengali and Hindi. *International Journal of Recent Trends in Engineering, 1*(1), 408-412.

Jain, A., Yadav, D., & Tayal, D. K. (2014, September). NER for Hindi language using association rules. In *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)* (pp. 1-5). New Delhi, India.

Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C., & Xu, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, *18*(5), 601-606.

Kazama, J. I., & Torisawa, K. (2008). Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of ACL-08: HLT* (pp. 407-415). Columbus, Ohio, USA.

Khalid, M., Jijkoun, V., & De Rijke, M. (2008). The impact of named entity normalization on information retrieval for question answering. In *European Conference on Advances in Information Retrieval* (pp. 705-710). Berlin, Heidelberg: Springer.

Kim, J. H., Kang, I. H., & Choi, K. S. (2002, August). Unsupervised named entity classification models and their ensembles. In *Proceedings of the 19th International Conference on Computational Linguistics-Volume 1* (pp. 1-7). Association for Computational Linguistics, Taipei, Taiwan.

Krishnarao, A. A., Gahlot, H., Srinet, A., & Kushwaha, D. S. (2009, May). A comparison of performance of sequential learning algorithms on the task of named entity recognition for Indian languages. In *Proceedings of the International Conference on Computational Science* (pp. 123-132). Berlin, Heidelberg: Springer.

Konkol, M., Brychcín, T., & Konopík, M. (2015). Latent semantics in named entity recognition. *Expert Systems with Applications*, *42*(7), 3470-3479.

Kozareva, Z. (2006, April). Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 15-21). Association for Computational Linguistics, Trento, Italy.

Li, W., & McCallum, A. (2003). Rapid development of Hindi named entity recognition using conditional random fields and feature induction. *ACM Transactions on Asian Language Information Processing (TALIP)*, *2*(3), 290-294.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203-208.

Lund, K., Burgess, C., & Audet, C. (1996). Dissociating semantic and associative word relationships using high-dimensional semantic space. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 603-608). Erlbaum, New Jersey.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, *35*(5), 482-489.

Moro, A., Raganato, A., & Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, *2*, 231-244.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3-26.

Patil, N., Patil, A. S., & Pawar, B. V. (2016). Survey of named entity recognition systems with respect to Indian and foreign languages. *International Journal of Computer Applications*, *134*(16), 21-26.

Pinto, D., McCallum, A., Wei, X., & Croft, W. B. (2003). Table extraction using conditional random fields. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 235-242). Association for Computing Machinery, Toronto, Canada.

Avinesh, P., & Karthik, G. (2007). Part-of-speech tagging and chunking using conditional random fields and transformation-based learning. In *Proceedings of the IJCAI Workshop on Shallow Parsing for South Asian Languages* (pp. 21-24). Hyderabad, Inda.

Rocktäschel, T., Weidlich, M., & Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, *28*(12), 1633-1640.

Rodriquez, K. J., Bryant, M., Blanke, T., & Luszczynska, M. (2012). Comparison of named entity recognition tools for raw OCR text. In *Proceedings of the 11th Conference on Natural Language Processing (KONVENS)* (pp. 410-414). Vienna, Austria.

Saha, S. K., Sarkar, S., & Mitra, P. (2008, January). A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I* (pp. 343-349). Hyderabad, India.

Saha, S. K., Sarkar, S., & Mitra, P. (2009). Feature selection techniques for maximum entropy based biomedical named entity recognition. *Journal of Biomedical Informatics, 42*(5), 905-911.

Saha, S. K., Mitra, P., & Sarkar, S. (2012). A comparative study on feature reduction approaches in Hindi and Bengali named entity recognition. *Knowledge-Based Systems*, *27*, 322-332.

Saha, S. K., Narayan, S., Sarkar, S., & Mitra, P. (2010). A composite kernel for named entity recognition. *Pattern Recognition Letters*, *31*(12), 1591-1597.

Sahin, H. B., Tirkaz, C., Yildiz, E., Eren, M. T., & Sonmez, O. (2017). Automatically annotated Turkish corpus for named entity recognition and text categorization using large-scale gazetteers. *arXiv preprint arXiv:1702.02363*.

Sha, F., & Pereira, F. (2003, May). Shallow parsing with conditional random fields. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 134-141). Association for Computational Linguistics, Edmonton, Canada.

Srivastava, S., Sanglikar, M., & Kothari, D. C. (2011). Named entity recognition system for Hindi language: a hybrid approach. *International Journal of Computational Linguistics (IJCL)*, *2*(1), 10-23.

Szarvas, G., Farkas, R., & Kocsor, A. (2006, October). A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. In *Proceedings of the International Conference on Discovery Science* (pp. 267-278). Berlin, Heidelberg: Springer.

Tayal, D. K., Ahuja, L., & Chhabra, S. (2015). Word sense disambiguation in Hindi language using hyperspace analogue to language and fuzzy c-means clustering. In *Proceedings of the 12th International Conference on Natural Language Processing* (pp. 49-58). Trivandrum, India.

Zhou, G., & Su, J. (2002). Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 473-480). Association for Computational Linguistics, Philadelphia, Pennsylvania.