A survey on the state-of-the-art machine learning models in the context of NLP

Wahab Khan[1,*], Ali Daud[2,1], Jamal A. Nasir[1], Tehmina Amjad[1]

[1]Dept. of Computer Science and Software Engineering, IIU, Islamabad 44000, Pakistan

[2]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

[*]Corresponding author: wahab.phdcs72@iiu.edu.pk

Abstract

Machine learning and Statistical techniques are powerful analysis tools yet to be incorporated in the new multidisciplinary field diversely termed as natural language processing (NLP) or computational linguistic. The linguistic knowledge may be ambiguous or contains ambiguity; therefore, various NLP tasks are carried out in order to resolve the ambiguity in speech and language processing.The current prevailing techniques for addressing various NLP tasks as a supervised learning are hidden Markov models (HMM), conditional random field (CRF), maximum entropy models (MaxEnt), support vector machines (SVM), Naïve Bays, and deep learning (DL).The goal of this survey paper is to highlight ambiguity in speech and language processing, to provide brief overview of basic categories of linguistic knowledge, to discuss different existing machine learning models and their classification into different categories and finally to provide a comprehensive review of different state of the art machine learning models with the goal that new researchers look into these techniques and depending on these, develops advance techniques. In this survey we reviewed how avant-grademachine learning models can help in this dilemma.

Keywords: Ambiguity; linguistic knowledge; machine learning; NLP; supervised learning.

1. Introduction

The recent years have witnessed a surge of interests in knowledge discovery from natural languages through machine learning techniques. The intent of natural language processing (NLP) or computational linguistic area is to study algorithms and methods for building computational models that are able to analyze natural languages for performing useful tasks like enabling communication between humans and machines, improving communication among humans or simply doing processing of text or speech (Jurafsky & James, 2000). The linguistic knowledge may be ambiguous or contains ambiguity. In order to resolve ambiguity in the discovered linguistic knowledge various NLP tasks e.g. POS, NER, SBD, word sense disambiguation and word segmentation are carried out using machine learning models. Machine learning models are decisive for resolving ambiguity as well as capturing every kind of linguistic knowledge (Jurafsky & James, 2000). Avant-garde NLP algorithms that are proposed in literature, depend on statistical machine learning or exclusively on supervised machine learning models. Before ML techniques, all NLP tasks are carried out using various rules based approaches, where sizably voluminous sets of rules are coded manually. The paradigm of machine learning is different from that of most prior endeavors at language processing. In literature, implementation of various ML techniques for various NLP tasks have been investigated extensively. These machine learning techniques may use parametric, non-parametric or kernel based learning algorithms. In ML based approaches, ML algorithms are trained in training phase on enough pre tagged data to generate model data, after that the model data are used in testing phase to test new data. Progressively, however, research has centered on stochastic machine learning models. In such model, to each input feature a real valued weight is attached, which generate soft probabilistic decisions. The benefit of such models is that: these models have the capability to represent a relation quality in different dimensions.

As per our knowledge to date, no other such a comprehensive survey paper as ours is reported in the field of machine learning and NLP. The work most relevant to ours is Nadeau & Sekine (2007), in which the authors just only provided an overview of the major techniques used for only name entity recognition (NER) task.

Our contributions in this work are as follows:

- To highlight various ambiguity types in speech and language processing

- To provide brief overview of basic categories of linguistic knowledge
- To discuss different existing machine learning models and their classification into different categories
- To provide a comprehensive review of different state-of-the art supervised machine learning models, which are addressed in literature for five major NLP tasks e.g. part of speech (POS) tagging, named entity recognition (NER), sentence boundary detection (SBD) and word segmentation.
- To provide brief description and to highlight construction methodologies of most commonly used dataset for major NLP

The structure of the survey is organized as follows: In Section 2 linguistic knowledge ambiguity is discussed. Section 3 describes basic categories of linguistic knowledge. Section 4 highlights different existing machine learning models and their classification into different categories. Section 5 provides a wider-angel review of different state of the art machine learning models. In section 6, dataset are explored. Section 7 highlights future direction and in section 8 conclusions is provided.

2. Ambiguity in terms of NLP

When in single passage for a single word or token there exist two or more possible meanings than this is termed as an ambiguity. The input text is said to be ambiguous, if multiple alternative linguistic structures can be built for it. Most of the NLP tasks e.g. word sense disambiguation, POS and natural language understanding or discourse analysis can be viewed as resolving ambiguity in the different categories of linguistic knowledge.

2.1. Lexical ambiguity

The lexical ambiguity is a kind of ambiguity in which a token or a sequence of tokens having different meanings in different contexts. In such cases a single word might have different meanings in the language to which it belongs. For flesh out, the word “bank” has several discrete lexical definitions; including “financial institution” and “edge of a river” similarly the word “saw” is used in three discrete senses in the preceding sentence: “I saw a saw, which could not saw”. In this sentence the word ‘saw’ has been used in three different meanings. Firstly the word ‘saw’ refer to a verb, secondly it refer to a tool name or noun and thirdly it again refer to a verb. So the word ‘saw’ is morphologically and syntactically ambiguous: noun or verb. Similarly the words ‘Pound’ and ‘Bat’ also create lexical ambiguity. The ‘pound’ might be weight, or they might be English money; similarly ‘bat’ might be flying mammal or a wooden equipment used by batsmen in the sport of cricket to hit the ball. The word ‘Cricket’ can be a name of an insect and also a name of a sport game. Lexical ambiguity can be addressed with tasks referred to as word sense disambiguation and POS.

2.2. Syntactic ambiguity

The main reason due to which a syntactic ambiguity arises is the structure of a sentence. It is the sentence structure due to which a sentence can have two or more than two meanings. The phenomenon of syntactic ambiguity often occurs, when adding an expression, such as a function word expression, the use of which is ill-defined. “He ate the cookies on the couch”, for example, one possible meaning of the said sentence can be: that those cookies that were on the couch was eaten by him and the second possible meaning can be that during sitting on the couch the person ate those cookies. Another example of Syntactic ambiguity is “Did you see the boy with the camera?” This question has two clear possible interpretations. This ambiguity arises from the prepositional phrase “with the Camera”. The two possible interpretations are as follows: “Did you see the boy, who is holding the camera?”

“Did you see the boy by using the camera?”

NLP tasks such as part of speech (POS) tagging, natural language understanding or discourse analysis can be used to resolve syntactic ambiguity.

2.3. Semantic ambiguity

In general, resolving semantic ambiguity in a plain text is recognized as word sense disambiguation. Also this category of ambiguity is more challenging compared to syntactical disambiguation. Semantic ambiguity takes place in situations when in a sentences there exists equivocal words or sequences of words that have multiple related meaning. For example, “We saw her duck”; here the word “duck” can refer to the girl’s bird or to a movement she made. Consider, in light of Facebook groups, the most commonly used statement: “Respected members, please inbox me the following articles, thanks”. So here in the statement the word ‘inbox’ is not clear. The term ‘inbox’ can be used for Yahoo mail, Gmail as well as for Facebook accounts. So, here from inbox, it is not clear which inbox the group member means. Consider another statement: “Would you like to join us in cup of coffee?” In this statement the proposition ‘IN’ create semantic ambiguity so the correct one is “Would you like to join us for a cup of coffee?” Semantic ambiguity can be resolved with tasks such as word sense disambiguation.

3. Linguistic knowledge concepts and terminology

3.1. Phonetics & Phonology

Both phonetics and phonology are described as study of speech and sounds. Phonetics is concerned with auditory perception and acoustic properties of speech. Phonology is related with phonemes, where phonemes are intellectual expressive units of speech, such that different languages have different phonemes. In order to make words, phonology also deals with the rules by which these sounds are constrained. For example, consider the morpheme ‘Subtle’, During pronunciation, the alphabet ‘b’ in word ‘Subtle’ is suppressed and pronounced as ‘Sutle’. Phonetics, however, is related to allophones, where allophones are the actual physical parts of various speech sounds. Basically, phonetics is concerned with the phenomenon that how various sounds are generated in the human vocal tract, how these sounds are transmitted by sounds waves and how our auditory system perceive these sounds. For example the difference between the verbalization of ‘t’and ‘d’, where the ‘t’ is voiceless and the’d’ is voiced. The core difference between phonetics and phonology is that, phonetics mainly deals with the description of speech sounds, whereas phonology deals with meaning. In short knowledge about linguistic sounds is called phonetics and phonology. This category of linguistic knowledge is more prone to lexical ambiguity.

3.2. Morphology

The terms morphology and syllable structure are commonly used interchangeably to each other in the literature of linguistics. In morphology linguistic basic units, such as root words, part of speech, affixes etc as well as morphemes of a given language are identified, analyzed in detail and structure is described deeply. Morphology is also known as the knowledge of meaningful component of words (Jurafsky & James, 2000). E.g. the morphology of English sentence is Subject (S), Verb (V) and Object (O) SVO, while the morphology of Urdu Language is SOV (Daud *et al.,*2016). Similarly, from morpheme ‘drink’ the word drinking (drink +ing) and so on. The most common type of ambiguity that arises in this category of linguistic knowledge is syntactic ambiguity.

3.3. Syntax

In linguistics, syntax is “the study of the principles and processes by which sentences are constructed in particular languages” (Tałasiewicz, 2009). Syntax is concerned with rules and codes that are necessary for sentence formation of any language of the globe. E.g. the sentence “Colorless green ideas sleep furiously” is syntactically right because it follows English descriptive linguistics rules. Syntactic ambiguity is the most common type of ambiguity that occurs in the syntax of any language.

3.4. Semantic

Semantics is concerned with meanings of a particular word/phrase in a sentence. In semantic the main focus is on the relationship formation between various tokens/word such as phrases, region, and symbols, and also explore that what these concepts stands for (Danker, 2000). E.g. the sentence “Colorless green ideas sleep furiously” is syntactically precise because it follows English grammar rules; but this sentence is semantically incorrect because it contains several contradictions, colorless things cannot be green. Other forms of semantics include the semantics of programming languages, formal logics, and semiotics (Tałasiewicz, 2009). In semantics, semantic ambiguity more often occurs.

3.5. Pragmatics

Pragmatics is a subfield of linguistics, which studies the slipway in which context change to content. Pragmatics encompasses speech act theory, talk in interaction and other approaches to language behavior in liberal arts, social science, linguistics and anthropology (Bygate *et al.,* 2013). Unlike linguistics, which analyze idea that is conventional or “coded” in a given language, linguistics studies how the communicating of meaning depends on geophysics and subject area psychological feature (e.g., descriptive linguistics, cognition, etc.) of the electro-acoustic electrical device and listener, and also on the linguistic environment of the auditory communication, any pre-existing knowledge about those involved, the inferred intent of the speaker, and other divisor. In this respect, pragmatics explicate how language exploiters are able to overtake patent expression, since meaning relies on the manner, place, moment etc. of an utterance. E.g. “I” just met the old tribe man and his son, coming out of the mosque.’ Or “I wouldn’t have thought there was room for the two of them”. The pragmatic ambiguity type occurs in pragmatic kind of linguistic knowledge.

4. Machine learning & NLP

State of the art NLP approaches generally adopt machine learning algorithms or more generally based on statistical machine learning. The prior attempts for linguistic processing were based on rules;, rules are synthesized to carry out different NLP tasks. Unlike rule based approach, ML approach automatically induces rules from training data. Machine learning algorithms usually consisted of intelligent modules which have capability to learn from historical data. Before ML approaches, NLP tasks are commonly carried out using rule based approaches. In Rule based approaches rules were constructed manually by linguistic experts or grammarians for particular tasks.

Now a days various ML algorithms are in use for carrying various NLP tasks. The parameters of these algorithms is historical data from which features are synthesized, and this features based data latterly used in prediction or classification. Progressively, however, the avant grade research in computational linguistic domain is based on machine learning models. In such a model, a real valued weight is attached to each input feature, which generate soft, probabilistic decisions. The benefit of such models is that: these models have the capability to represent a relation quality in different dimensions.

In order to carry out major NLP task using statistical approaches, it incorporates stochastic and probabilistic methods.The use of ML algorithms is not limited only to NLP domain, but its application can also be found in health care domain and air pollution. Moses (2015) presented a comprehensive survey of different datamining algorithms used in cardiovascular disease diagnosis. Barakat *et al.,* (2014) used sub-sampled bootstrapping method a semi supervised machine learning model for air pollution modeling.

The current state of the art techniques that are widely used for major NLP tasks are machine learning (ML) techniques. Broadly Machine learning (ML) techniques can be put in three categories: (1) supervised learning, (2) semi-supervised learning and (3) unsupervised learning (Daud *et al.,* 2016).

The current dominant technique for addressing problem in NLP is supervised learning. The basic idea behind supervised machine learning models is that it automatically induces rules from training data. Supervised learning can be (a) sequential and (b) non-sequential.

Sequential supervised machine learning techniques are hidden Markov model (HMM), conditional random fields (CRF), maximum entropy (MaxEnt), and deep learning (DL), while non-sequential supervised machine learning techniques includes support vector machines (SVM), decision trees (DT) and Naïve Bays.

Semi supervised machine learning technique involve small degree of supervision, example is bootstrapping.

In unsupervised machine learning, the model is not trained. The task is achieved by finding intra- similarity and inter-similarity between objects. The most common approach of unsupervised category is clustering.

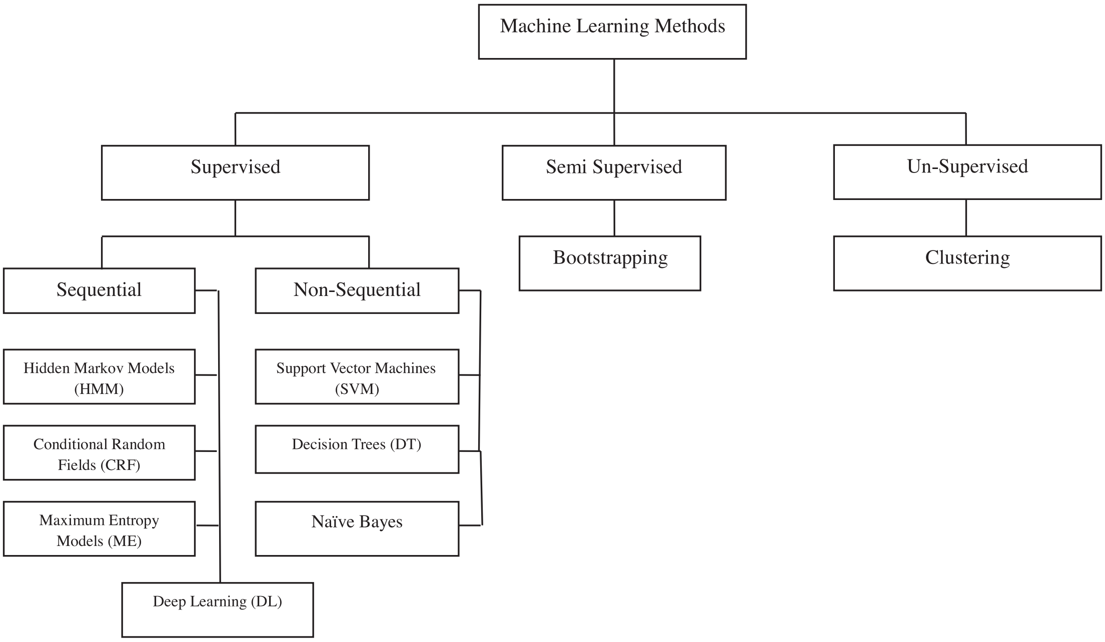

Fig. 1. State of the art machine learning models

5. Models

The current dominant technique for addressing problems in NLP is supervised learning. The basic idea behind supervised machine learning models is that it automatically induces rules from training data. The most frequent machine learning models that are commonly used for ambiguity resolution in the linguistic knowledge with major NLP tasks are: hidden Markov model (HMM), conditional random fields (CRF), maximum entropy (MaxEnt), support vector machines (SVM), decision trees (DT), Naïve Bays, and deep learning.

5.1. Hidden Markov model description and applications

Hidden Markov model is also termed a sequence classifier or sequence labeler. The basic function of a sequence classifier is: for given input sequence first to identify class label for each token and then to assign the corresponding label to each token or word of the sequence. Generally there are two types of sequence classifiers (1) non probabilistic sequence classifier e.g. finite state automata and (2) probabilistic sequence classifier e.g., HMM and ME. The parameters of HMM mentioned in Youzhi, (2009) are emission probabilities or observation probabilities, transition probabilities of states and a group of symbols. Further on the basis of random function parameter hidden Markov model can be classified in three categories. These categories are: (a) discrete hidden Markov model (b) continuous hidden Markov model and (c) semi-continuous hidden Markov model. Discrete hidden Markov model is referred as DHMM, and key logic behind this type its dependency on discrete probability density function, continuous hidden Markov model are also referred as CHMM, and its density function is based on continuous probability while the third type semi-continuous hidden Markov model referred to as SCHMM and consider good feature of both CHMM and DHMM for its density function.

Generally it is observed that in presence of right and enough training data, CHMM usually outperform DHMM and SCHMM (Youzhi, 2009). The hidden Markov model (HMM) can be considered as a probabilistic generative model of a sequence (Todorovic *et al.,* 2008). The following equation is used to express Hidden Markov model:

$$M = (A, B, \pi) \tag{1}$$

Where '**A**' represents matrix of transition probabilities:

$$A = (a_{ij}) \tag{2}$$

$$a_{ij} = P(s_i|s_j) \tag{3}$$

'**B**' represents matrix of observation probabilities:

$$B = (b_i(v_m)) \tag{4}$$

$$b_i(v_m) = P(v_m|s_i) \tag{5}$$

'**π**' Represents a vector of initial probabilities:

$$\pi = (\pi_i) \tag{6}$$

$$\pi_i = \mathrm{P}\ (s_i) \tag{7}$$

Hidden Markov model instantiates two assumptions (Jurafsky & James, 2000). Firstly, in any particular state probability calculation only the probability of previous state is considered.

Markov assumption:

$$P(q_i|q_1 \dots q_{i-1} = P(q_i|q_{i-1}) \tag{8}$$

Second, the probability of the resultant observation O_i make use of the probability of current state q_i which generates observation, and it is independent of other states around it:

Output independent assumption:

$$(o_i|q_1 \dots q_i, \dots, q_T, o_1, \dots, o_i, \dots o_T) = P(o_i|q_i) \tag{9}$$

5.1.1. Named entity recognition (NER).

In computational linguistic pipe line named entity recognition or identification is a prominent task. The major application areas in which NER can be incorporated includes: question answering, information extraction, machine translation etc. In named entity recognition (NER) task, the focus is to find out person name, location names, brand names, abbreviations, designation, date, time, number etc. and classifying them into predefined different categories (Singh *et al.,* 2012). The most recent work for NER task using HMM are Morwal & Chopra (2013) and Morwal & Jahan (2013). Morwal & Chopra (2013) have developed a tool named NERHMM for the task of named entity recognition based on hidden Markov model. The input to their NERHMM tool is pre labeled data from which the tool generates parameter of hidden Markov model e.g. (a) start probability (b) transition probability and (c) emission probability. The authors proposed NER HMM based model operates on sentence wise data and as a results assign correspond NE tag to each word in sentence. The primary objective of the authors was to develop NER toll for only Indian language but latterly when tested for other Indian languages such as Hindi, Bengali, Urdu, English, Punjabi and Telugu, promising results were observed. Morwal & Jahan (2013) has used hidden Markov model based machine learning approach for named entity recognition in Indian languages e.g. Hindi, Marathi and Urdu languages. To perform NER task in Hindi, they have used tourism domain corpus and for Marathi language NLTK Indian corpora is considered. So far the corpus for Urdu language is concerned; a corpus is created by translation of Hindi language corpus into Urdu language corpus by using Google translator. They have trained and tested their proposed HMM based model for NER task on 100 sentences and the tags are PER, LOC, COUNTRY, STATE, CITY, MONTH and OTHER. For Hindi language 86% accuracy, for Marathi 76% and for Urdu language 65% accuracy were recorded.

Zhou & Su (2002) have proposed a name entity recognition system based on HMM and HMM-based chunk tagger to predict and classify named entities such as person names, times and numerical units. The proposed HMM is based on the mutual information independence assumption, instead of the conditional probability independence assumption. The goal of Zhou & Su (2002) HMM based model was to straightforwardly produce the definitive NE tags from the yield expressions of the boisterous channel. The proposed HMM based chunk tagger is different from traditional HMM tagger in two respects; (a) it works in reverse order as compared to the traditional HMM and (b) the author proposed model consider mutual information independence while the classical HMM is based on conditional probability independence. A HMM-based NER framework has been accounted in Ekbal *et al.* (2007), where more context oriented data has been acknowledged throughout the emanation probabilities and NE additions have been kept for taking care of the obscure words. The model presented by Ekbal *et al.* (2007), is based on first Markov assumption for a particular NE tag probability calculation. For Markov, first assumption they have used trigram model, where the probability of particular NE tag totally dependent on n-2 previous tags instead of n-1 previous tag. To represent the beginning of sentence, the authors have also used an additional tag '$', which they termed a 'Dummy Tag'. For the issue of inadequate information, a linear interpolation technique has been utilized to smoothen the trigram probabilities.

5.1.2. Parts of speech (POS) tagging

Parts of speech are a well-known assignment in common dialect preparing provisions, which assume a paramount part in different requisitions like discourse distinguishment, data extraction, content to-discourse and machine interpretation frameworks (Anwar *et al.,* 2007). The POS tagging process was characterized by the author in Jurafsky & James (2000) as: POS tagging is the procedure by which a particular tag is allotted to each one expression of a sentence to demonstrate the capacity of that statement in the particular setting. Parts of speech tagging is basically a sequence classification task, where each one expression in a succession must be allocated a grammatical form tag. Youzhi (2009) is of the view that though Morphological analysis (MA) and part of speech (POS) tagging are two separate and independent problems of English, but as a research issue they are dependent on each other. To handle the morphological analysis (MA) issue, the author first used a knowledge-based techniques and for POS tagging they have used hybrid approach of rule-based method along with hidden Markov model (HMM). Elhadj (2009) developed a tagger for Holy Quran. Their developed tagger is based on an approach, which make use of both morphological examinations with hidden Markov models (HMMs). In their proposed approach Arabic sentence structure plays a vital role. They have utilized morphological examination to lessen the span of the tags vocabulary by sectioning Arabic token into its corresponding basic units such as: prefixes, stems while hidden Markov models (HMMs) is utilized to speak the Arabic sentence structure keeping in mind the end goal to consider the etymological consolidations.

5.1.3. Sentence boundary detection (SBD)

Sentence boundary detection is a preparatory venture for arranging a content archive for natural language processing errands, e.g., machine interpretation, POS tagging, content outline and so forth. The authors Liu, *et al.* (2006) and Rehman & Anwar (2012) have used HMM to detect sentence boundaries or punctuation in speech. The target of the authors was to advance such a framework, to the point that immediately includes data about the area of sentence limits and discourse disfluencies with a specific end goal to improve discourse distinguishment yield.

In Kolár & Liu (2010), the focus of author was to incorporate HMM for disambiguation sentence boundary detection task in speech automatically. Their HMM based proposed approach makes use of both textual and prosodic information. In their proposed methodology, both textual and prosodic information are used to find out the locality position of sentence-like unit (SU). The method proposed in Kolár & Liu (2010) to handle sentence boundary detection problem automatically in speech makes use of independent language and prosody model. The authors used trigram LMs with modified Kneser-Ney smoothing. They incorporated a decision tree classifier on prosodic model in order to obtain the observation likelihood. For data skewness problem and for decreasing classifier variance, a combination of ensemble sampling with bagging has been employed. A sentence boundary detector system: Bondec was introduced in Wang & Huang (2003). In Bondec for the task of SBD, the authors used HMM, rule based and maximum entropy approach. The three models are totally independent from each other in functionality.

5.1.4. Word segmentation

Hidden Markov models (HMM) are intensively adopted for numerous NLP tasks including word segmentation problem. The authors Gouda & Rashwan (2004) have used discrete hidden Markov models for segmentation of Arabic words into letters. In Wenchao *et al.* (2010) the authors compared in detail different machine learning models for Chinese word segmentation task including HMM. Primarily for models precision and efficiency evaluation, different tag sets are chosen. Secondly they compared tradition HMM with MEMM by suppling similar features to both models. They also compared the two model by providing different features to each model, to test individual feature impact on Chinese word segmentation. The accuracy reported by their HMM based model for word segmentation task was 91.15% and determines 96.48% of all the word boundaries correctly. The main advantage of the authors approach is that: (a) it does not need a large lexicon of Japanese words, (b) avoids knowledge-based or rule based methods.

5.2. Conditional random field description and applications

Conditional random field was introduced by Lafferty *et al.* (2001) as a statistical modeling tool for pattern recognition and machine learning using structured prediction.

Conditional random fields are a statistical models and was initially proposed by Lafferty *et al.* (2001) for solving pattern recognition task by making use of structured probabilistic prediction. Now a day, in NLP domain CRFs are the most widely adoptable statistical models for solving wide range of NLP tasks. CRFs have the advantage that it makes use of good features of both discriminative models as well as undirected graphical models. The major domains in which these models are widely used for prediction and classification tasks are: natural language processing, bioinformatics and computer vision. Conditional random fields models are based on conditional distribution instead of joint distribution of a set of response instances (Yang *et al.,* 2013). The major types of this class of model are categorical-discrete CRFs, skip chain CRFs, and conditional Gaussian based CRFs (Yang *et al.,* 2013). The application of CRF is not limited to NLP, but also have wide range of application in medical, engineering, energy forecasting etc.

Lafferty *et al.* (2001) defined CRF on observations X and random variables Y as follow:

Let G= (V, E) be a graph such that, Y= (Yv) v€V so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field when the random variables Yv, conditioned on X, obey the Markov property with respect to the graph:

$$P(Y_v|X, Yw, W \neq v) = P(Yv|X, Yw, w \sim V) \qquad (10)$$

Where w ˜ V means that w and v are neighbors in G.

CRF model aims at finding the label y, which maximizes the conditional probability p (y | x) for a sequence x. CRF models are a feature-based models and ubiquitously operable on both binary as well as real valued features.

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{t=1}^{T} \lambda_k f_k(y_t, y_{t-1}, x, t)\right) \qquad (11)$$

Where Z is the normalization factor, $f_k(y_t, y_{t-1}, x, t)$ is a feature function and λ_k is the learned weight for each feature function.

5.2.1. Named entity recognition (NER)

The state of the art approach that is currently used for NER is supervised learning approach. The most recent CRF based work for NER task is proposed in Liu *et al.* (2013). The authors have proposed a unique method of NER for tweets based on hybridization of a K-Nearest neighbors (KNN) classifier with a conventional linear conditional random fields (CRF) model. The authors proposed method conducts tweet normalization and combines a KNN classifier with a conventional CRF-based labeler under a semi supervised learning framework, to combat the lack of information in a single tweet as well as the unavailability of training data. A CRF model with bootstrapping technique for Arabic named entity recognition was incorporated in Abdel Rahman *et al.* (2010). The authors applied their proposed integrated approach to handle named entity recognition problem in Arabic language. The NE tag set used in their experiment was consist 10 NE's classes. The work of authors in Yao *et al.* (2009) tries to tackle NER problem in Chinese language. The author proposed model for Chinese NER task makes use of CRF, which was based on pool-based active learning algorithm. Moreover a CRF based Chinese NER system has been reported in Zhang *et al.* (2008), where fusion of multiple features are used to accomplish Chinese NER task robustly and accurately. In Benajiba & Rosso (2008) the authors have tried to enhance their previous work on Arabic named entity recognition, their new approach is the hybridization of maximum entropy and CRF, and results show that their hybridized approach out performs their previous work.

5.2.2. Parts of speech (POS) tagging

Recently Ammar *et al.* (2014) proposed joint use of CRF with auto-encoder technique for unsupervised learning. For prediction task authors trained CRF using rich features. The authors evaluated their proposed model performance on part of speech as well as on bi-text word alignment task. Pandian & Geetha (2009) proposed a language model for POS and chunking tasks based on CRF model for Tamil language. Here the authors compared their proposed CRF model with baseline CRF model for POS tagging and chunking, and showed that their model gives high performance as compared to baseline CRF model. Work on the POS tagging for Guajarati language has been reported by Patel & Gali (2008), that uses Conditional Random Fields (CRF) for tagging and chunking tasks. Their proposed model make use of a tag set of 26 POS tags, almost defined for the other Indian languages.

5.2.3. Sentence boundary detection (SBD)

Sentence boundary identificationis a preparatory venture for planning a content archive for natural language processing errands, e.g., machine interpretation, POS tagging, content rundown and so forth (Rehman & Anwar, 2012). CRF has been applied for the first time to sentence splitting and tokenization in scientific documents from the biomedical domain in Tomanek *et al.* (2007). In their proposed model, the task is completed in two phases. First they split the entire text document into constituent sentence and in second phase the sentences are split to constituent tokens based on CRF.

5.2.4. Word segmentation

So far to our knowledge, Peng *et al.* (2004) are the pioneers, who accomplished word segmentation task in Chinese language with the usage of conventional CRF. The task of word segmentation was treated as binary classification in Peng *et al.* (2004). The accuracy rate reported in Peng *et al.* (2004) for their Chinese word segmenter is 95% on UPenn Chinese Treebank dataset. The earlier implementations of CRFs based on assignment of conditional probability for a label sequence $X = x_1$………. x_T given an observation sequence $O = O_1$………..O_T. While in the author's formulation, CRFs deal with word boundary ambiguity. Lafferty *et al.* (2001) introduced conditional random fields (CRFs) to solve the label bias limitation of MaxEnt model for English word segmentation task in a principled way. The authors' CRF has a solitary exponential model for the joint likelihood of the whole succession of names subject to perception grouping.

5.3. Maximum entropy description and applications

Maximum entropy models offer a clean approach to join various bits of logical confirmation, keeping in mind the end goal to gauge the likelihood of a certain phonetic class happening with a certain semantic setting. Maximum entropy is a supervised probabilistic machine learning model used for sequential data classification. Probabilistic classifier is type of a classifier, where the classifier equally distributes probability over all classes for a given observation sequence, in addition to assigning a label or class.

Maximum entropy models complete its task in three steps; firstly from given input sequence it extract relevant features, secondly perform linear combination of the extracted features and finally taking exponent of resulted sum (Jurafsky & James, 2000). The probability distribution of a certain class 'x' given the observation '*o*' is given as:

$$P(x|o) = \frac{1}{Z}\exp\left(\sum_{i=0}^{n} w_i f_i\right) \tag{12}$$

Where, Z is a normalization function and *exp* = e^x

5.3.1. Named entity recognition

Benajib *et al.* (2007) presented ANERsys: purely Arabic text. The presented generic Arabic NER system was based on n-grams and maximum entropy techniques. In Benajiba *et al.* (2007) the authors improved their own particular preparing and test corpora (ANERcorp) and gazetteers (ANERgazet) to prepare, assess and help the actualized system. In Saha *et al.* (2008) the authors proposed Hindi NER system based on maximum entropy (MaxEnt). In first phase the authors focused on the identification of most feasible feature for Hindi NER task. The key feature on which they focused consisted of both lexical as well as context window features. Moreover they have also used other lexical resources such as gazetteer list for their model. In Borthwick (1999), the author has developed a novel system called "MENE" which stands for maximum entropy named entity for tagging named entities in text. In their proposed model, they have tried to model conditional probabilities instead of joint probabilities. Their proposed model first set up feature pool, in second step the corpus is tokenized for further process.

5.3.2. Parts of speech (POS) tagging

The research work of Ratnaparkhi (1996) breaks new ground for using machine learning models for major NLP tasks. Ratnaparkhi (1996) has reported first time ME model for parts of speech tagging task. The model presented by Ratnaparkhi (1996) trains from a corpus annotated with part-of-speech tags and assigns them to previously unseen text. Ekbal *et al.* (2008) have explored the use of ME for Bengali part of speech tagging task. The system developed by Ekbal *et al.* (2008) for Bengali part of speech tagging assignment make utilization of the distinctive logical data of the words in addition to the mixed bag of characteristics that are accommodating in foreseeing the different POS classes. Their POS tagger has demonstrated an accuracy of 88.2% for a test set of 20K word forms.

5.3.3. Sentence boundary detection (SBD)

So far to our knowledge, maximum entropy (MaxEnt) for sentence boundary detection is incorporated in Reynar & Ratnaparkhi (1997) for English language. The task of their presented trainable model, based on maximum entropy is to identify sentence boundaries in raw text. The training procedure of their presented model does not require any hand-crafted rules, part-of-speech tags, or domain-specific information. Agarwal *et al.,* (2005) proposed a maximum entropy based model for sentence boundary detection task. In experiment the authors make use of only context features of trigrams. They evaluated the performance of their proposed MaxEnt based model on dataset namely Wall Street Journal (WSJ), Penn Treebank and GENIA. The trigram context consists of current word, previous and next word along with its corresponding tag.

5.3.4. Word segmentation

In Xue (2003) the author has reported ME based approach for Chinese word segmentation task. The author trained his maximum entropy based tagger on manually annotated data to automatically assign to Chinese characters tags that indicate the position of a character within a word. The tagged output is then converted into segmented text for evaluation. A maximum entropy based model is presented in Luo (2003) for Chinese language, where Chinese character based parser is incorporated, which does word-segmentation, POS tagging and parsing in a unified framework. Low *et al.* (2005) evaluated their Chinese word segmenter based on maximum entropy model on four data sets namely Academia Sinica (AS), City University of Hong Kong (CITYU), Microsoft Research (MSR), and Peking University (PKU). On experimental basis they concluded that the segmentation accuracy rate of ME based Chinese word segmenter can be improved with the use of an external dictionary and additional training corpora. The major challenge to Chinese word segmenter, which can decrease the accuracy level mentioned in Low *et al*. (2005) is the use of out-of-vocabulary (OOV) words.

5.4. Support vector machine description and applications

Support vector machine (SVM), is a machine learning algorithm used for classification of both linear and nonlinear data, mainly for binary classification. It utilizes a nonlinear mapping to convert the definitive preparing information into a higher size. Inside this new measurement, it hunts down the direct optimal dividing hyper plane. The hyper plane can be used to separate the data of two classes. In the field of natural language processing, SVMs are applied to number of NLP tasks e.g. POS, NER, segmentation, content arrangement and so forth, and are accounted to have attained high exactness without falling into overfitting in spite of the fact that with an expansive number of words taken as the characteristics (Ekbal & Bandyopadhyay, 2008). So far the performance of SVM is concerned, it produces highly accurate results, but the training time is extremely slow. The classification rule for separating hyperplane can be written as:

$$f(x, w, b) = W.X + b = 0 \tag{13}$$

Where W is a weight vector, namely, W= w_1, w_2, w_3, w_4 …n, n is the number of attributes; x the example to be classified and b is a scalar, often referred to as a bias. The above equation can be written as:

$$w_0 + w_1x_1 + w_2x_2 = 0 \tag{14}$$

Where x_1 and x_2 are the values of attributes A_1 and A_2, respectively, for X and b as an additional weight, w_0.

Thus, any point that lies above the separating hyperplane satisfies the below condition.

$$w_0 + w_1x_1 + w_2x_2 > 0 \tag{15}$$

Similarly, any point that lies below the separating hyper plane satisfies:

$$w_0 + w_1x_1 + w_2x_2 < 0 \tag{16}$$

5.4.1. Name entity recognition

In Ekbal & Bandyopadhyay (2009), multi engine system is used based on the combination of SVM, CRF, ME for Bengali named entity recognition. The training corpus consists of 272K word forms, out of which 150K word forms have been manually annotated with the four major named entity (NE) tags, namely Person name, Location name, Organization name and Miscellaneous name. Their Comparative evaluation results also show that the proposed SVM based system outperforms the three other existing Bengali NER systems. NE recognizer based on support vector machines (SVMs) gives better results than conventional systems (Isozaki & Kazawa, 2002). The main objective of the authors was to present a method that makes the NE system substantially faster as SVM classifiers are too inefficient for NE recognition. Their SVM-based NE recognizer attained accuracy rate of 90.03%. The improved classifier is 21 times faster than TinySVM and 102 times faster than SVM-Light.

5.4.2. Part of speech (POS) tagging

POS tagging is a very important preprocessing task for language processing activities. SVM is used for Bengali part of speech task by Ekbal & Bandyopadhyay (2008). The POS tagger developed by Ekbal & Bandyopadhyay, 2008) makes use of 26 POS tag set defined for Indian languages. A POS tagger for Malayalam language was built using support vector machine (SVM) by Antony *et al.* (2010). The authors first identified the ambiguities in Malayalam lexical items, and developed a tag set consisting of 29 tags, which was appropriate for Malayalam. Their SVM model receives the corpus data in tokenized form. Their proposed architecture consists of five steps namely tokenization, manual tagging, corpus training, tagging using SVM and SVMT. Sajjad & Schmid (2009) are the first ones who introduced the use of SVM model for Urdu language part of speech tagging task. In their research work they compared the results of SVM model with TnT tagger, Tree tagger, RF tagger and experimentally showed that SVM tool shows the best accuracy of 94.15%.

5.4.3. Sentence boundary detection

Gillick (2009) described a simple yet powerful method for Sentence boundary detection (SBD) task using support vector machines. In their work they discussed the main reason, which makes the sentence boundary task challenging, which feature are relevant to be considered and developed a statistical system based on SVM model for sentence boundary detection task. In Akita *et al.* (2006) the author introduced SVM model for sentence boundary identification of spontaneous Japanese. The creators received SMV model to acknowledge vigorous classification against a wide mixed bag of articulations and discourse distinguishment mistakes. Recognition is performed by a SVM based content chunker utilizing lexical and stop data as characteristics. In their study, the authors compared the results generated by SVM based text chunker with statistical language model (SLM) and concluded that SVM provide high accuracy.

5.4.4. Word segmentation

Haruechaiyasak *et al.* (2008) analyzed and compared Naive Bayes (NB), decision tree, support vector machine (SVM), and conditional random field (CRF) approaches for Thai word segmentation. Their results show that CRF provide better results than that of the others ML techniques. In Nguyen *et al.* (2006) the authors report a careful investigation of using conditional random fields (CRFs) and support vector machines (SVMs) for Vietnamese word segmentation. They trained both models with different feature settings. SVMs are binary classifiers, and are extended to multi-class classifiers in order to classify three or more classes (Nguyen *et al.,* 2006).

5.5. Naïve Bays description and applications

Naïve Bayes classifiers are statistical classifiers. They can anticipate class enrollment probabilities; for example, the likelihood that a given specimen has a place with a specific class is dependent upon Bayes hypothesis, which is a straightforward and effective likelihood arrangement system dependent upon supervised classification technique (Han *et al.,* 2006). Naive Bayes classifier needs just little measure of preparing set to gauge the parameters for characterization (Sunny *et al.,* 2013). The classifier is stated as

$$P(H|X) = P(X|H) * P(H)|P(X) \tag{17}$$

Where P (H) is the prior probability of H, P (H|X) is the conditional probability of H, given X called the posterior probability, P (X|H) is the conditional probability of X given H and P (X) is the prior probability of X.

Naïve Bayes classifier is based on class conditional independence assumption. In class conditional independence assumption, the effect of an attribute value on a given class is independent of the value of the other attributes (Han *et al.,* 2006).

The Naïve Bayes classifier currently experiences a new beginning in the field of computational linguistics. The contribution of Naïve Bayes technique in computational linguistic is not promising. Only few research works so far reported on the use of Naïve Bayes technique for NLP tasks are Gillick (2009), Sunny *et al.* (2013) and Ahmed & Nürnberger (2009).

5.5.1. Named entity recognition (NER).

The main objective of proposed work in Mohit & Hwa (2005) was to study the role of syntactic features in building a semi supervised named entity (NE) tagger. For this purpose they trained a Naive Bayes classification model on a combination of labeled and unlabeled examples with the expectation maximization (EM) algorithm. The authors concluded that a significant improvement in classification accuracy can be achieved by combination of both dependency and constituency extraction methods. In Zhang *et al.* (2004), the author proposed a statistical model for focused or most topical named entity recognition by converting it into a classification problem. To address focused or most topical named entity recognition problem, they compared three classification methods: a decision tree based rule induction system, a Naive Bayes classifier, and a regularized linear classification method based on robust risk minimization. Finally, they demonstrated that the proposed method can achieve near human-level accuracy.

5.5.2. Word segmentation

In Zheng & Tian (2010), the proposed scheme is for developing a Naïve Bayes based model from Chinese web text categorization. The authors experimentally showed that the results of their developed system are more accurate and is more efficient as compared to previous ones. Their proposed system works in three phases; first: obtain training text set, second: establish text representation model and finally text feature extraction process. In Haruechaiyasak *et al.* (2008), the proposed work analyzed and compared various approaches to Thai word segmentation task. Among machine learning approaches, they also tested the performance of Naïve Bays model for Thai word segmentation task. They first compared dictionary based approach with machine learning based approach and showed that dictionary based approach outperform machine learning approach in terms of precision and recall. Among machine learning approach CRF model outperform the other three ML model including Naïve Bays model.

5.6. Deep learning description and applications

Deep learning algorithms are popular types of ML algorithms, which tries to learn from layered model of inputs, commonly known as neural nets. In deep learning approach concept, learning of a current layer is dependent on previous layer input. With each succeeding layer, deep learning algorithm tries to learn numerous levels of concept of increasing complexity/abstraction (Deng & Yu, 2014). Deep learning algorithms can fall in both supervised and unsupervised categories. The main applications of deep learning includes pattern recognition and statistical classification. Various deep learning architectures such as deep neural networks, convolutional deep neural networks, and deep belief networks have been successfully incorporated in various domain like computer vision, automatic speech recognition, natural language processing, audio recognition and bioinformatics, where they have been observed to produce auspicious results on different tasks. Notable results have been reported by applying deep neural algorithms in the field of NLP for the tasks of sentiment analysis, parsing, IR, NER and other areas of NLP.

5.6.1. Part of speech tagging

Zheng *et al.* (2013) incorporated deep learning approach for Chinese word segmentation and part of speech tasks. For relevant features discovery, they used deep learning algorithms. For the improvement of Chinese character representation, they make use of large amount of unlabeled Chinese data. They latterly used these improved Chinese character representations for segmentation and POS tasks improvement. The performance achieved by their proposed approach were closely related to the avant grade approaches. The main contribution of the author was (i) to describe a perceptron-style algorithm for training the neural networks and (ii) presents the general architecture of neural networks. Santos & Zadrozny, (2014) proposed new deep neural network (DNN) architecture that joins word level and character level representations to perform POS tagging. The proposed DNN, which they call CharWNN, uses a convolutional layer that allows effective feature extraction from words of any size. At tagging time, the convolutional layer generates character-level embedding for each word; even for the ones that are outside the vocabulary. The authors produced state of-the-art POS taggers for two languages: English, with 97.32% accuracy on the Penn Treebank WSJ corpus; and Portuguese, with 97.47% accuracy on the Mac-Morpho corpus.

5.6.2 Named entity recognition

Significant amount of work has been reported in literature to tackle major NLP tasks such as POS, word segmentation using deep learning approaches. On the other hand, NER lags far behind in terms of deep learning approach adaptation. Mohammed & Omar (2012) achieved Arabic NER task using neural network approach. The authors achieved Arabic NER task in three phases. In first stage, they preprocessed the corpora for onward processing. In stage two, they converted Arabic text to its roman equivalent and finally they applied neural network to achieve NER task. The authors compared their proposed approach with decision tree using the same data. The results achieved were 92 %, which show that the neural network outperform decision tree approach.

The authors achieved Arabic NER task in three phases. In first stage, they preprocessed the corpora for onward processing. In stage two, they converted Arabic text to its roman equivalent and finally they applied neural network to achieve NER task. The authors compared their proposed approach with decision tree using the same data. The results achieved were 92 %, which show that the neural network outperform decision tree approach.

5.6.3. Word segmentation

Chinese word segmentation task was investigated by Li *et al.* (2005) using perceptron based learning algorithm. They used four corpora in their experiment. The corpora includes: “As”, “PKU”, “CITYU” and “MSR”. They achieved word segmentation task by using character based classification system. To do so, they chunked original problem in various binary problems. After testing, results show that the proposed system performs well on three corpora (“AS”, “CITYU” and MSR”), while perform worse on the remaining ones. Results for each of the four corpuses were (F1%): 95.27, 95.14, 94.99 and 94.12. Qi *et al.* (2014) introduced deep neural network system for information extraction task. They tested the system on character-based sequences. Their character based sequences includes: Chinese NER and detection of secondary structure in protein sequence.

The proposed discriminative framework includes three important schemes. Firstly to capture semantic relationship between characters, they incorporated a deep learning based module, mapping characters to vector representations. Secondly, to improve vector representation, they adopted semi supervised learning by utilization of online sequences and finally they demonstrated spatial dependency constraint among labels. The authors performed experiment on CTB dataset for word segmentation task; and showed better performance. Word segmentation tagging results in improved F1 measure from 94.73% to 95.57%.

Table 1 provides summary of different supervised ML models including HMM, CRF, SVM, DT, maximum entropy and Naïve Bayes for five major NLP tasks e.g. NER, POS, sentence boundary detection, word segmentation and word sense disambiguation. From the table, it clear that NER and POS tasks are widely investigated with almost all supervised ML models, as compared to the other three NLP tasks. The tasks in the Indic languages are less addressed with ML techniques, as compared to its European counterpart. The NLP research community showed a lot of research interest in the exploration of NLP tasks with ML techniques in European languages, particularly in English language. The biggest reason for high research ratio in English language is the availability of lot of language resources. e.g. the availability of large corpora. So far, Naïve Bays model is concerned, the Naïve Bayes classifier currently experiences a new beginning in the field of computational linguistics and the contribution of Naïve Bayes model in addressing NLP task is not so promising.

Table 1. Summary of different machine learning models for various NLP tasks.

Lang\ Task	Name Entity Recognition (NER)		Parts of Speech Tagging (POS)		Sentence Boundary Detection (SB)		Word Segmentation (WS)	
	Model	**Year(s)**	**Model**	**Year**	**Model**	**Year**	**Model**	**Year**
Eng.	HMM	1998, 2002, 2008	HMM	2007, 2009	HMM	2004, 2006	CRF	2001
	CRF	2013	DT	1999	CRF	2007		
	ME	1999	ME	1996	ME	1997, 2005		
	DT	2006			DT	1998		
	DL	2013	DL	2001, 2014	SVM	2009		
Arab.	CRF	2010	HMM	2009			HMM	2004
	ME	2007					SVM	2005
	DL	2012						
Beng.	CRF	2008	CRF	2007				
	SVM		SVM	2008				
Hind.	CRF	2008					HMM	2008
	ME	2008						
Urdu.	CRF	2008	SVM	2010			HMM	2010
			CRF	2010				
			ME	2010				
Chin.	CRF	2008, 2009	HMM		CRF	2010	CRF	2004
	DL	2010	DL	2013	DT	2012	ME	2003, 2004, 2005
							Naïve Bays	2010
							DL	2014, 2008, 2005
Mani:	CRF	2011	CRF	2008				
Vitn.			ME	2010	ME	2008	SVM	2006
Myan .			CRF	2011				
Tamil.			CRF	2009				
Guaj.			CRF	2008				
Thai							DT	2000, 2001
							HMM	2009
							Naïve Bays	2008

6. Datasets

Table. 2 provides a brief summary of the most commonly used datasets for major NLP tasks carried out through supervised machine learning models, along with corresponding accuracy level of each model. Wall Street Journal, Penn-TreeBank and Brown Corpus have been widely utilized by researchers for experimentation. In Treebank dataset, English sentences are tagged with parts of discourse. The Brown Corpus, the most widely used corpus in current linguistic processing, actually contains more clean contents of pure American English and its size is around a million words from wide range of sources.

Most researchers have manually created their own dataset from web archives. Such datasets can be seen in Ekbal & Bandyopadhyay (2010). Daily paper on the web is an immense wellspring of promptly accessible dialect information. Most newspapers have their web version in the web and some of them furnish their chronicle accessible additionally. The documents are extracted from the web through web crawler and stored on a central location. After storing the contents in next step, the contents are refined so that it can be used as a corpus in future for various NLP tasks. The job of web crawler is to retrieves interested contents from online pages or archives and to store it in formats such as XML etc. Once the pages are retrieved, then the extracted HTML pages are cleaned. After cleaning the pages, a tag set is defined for annotation of the corpus. Web crawlers accomplish their task in steps. First it crawls a portion of a specified site, secondly it identify the interested contents e.g. the data, in third step index the data and finally perform search.

The focus of Antonova & Misyurev (2011) was to find out such a pool of techniques and procedure through which development of web-based parallel corpus for Russian and English languages can easily be accomplished. The resulting parallel corpus contains parallel sentences of both English and Russian languages. This corpus can be used to train ML techniques to achieve ML based translation task. The resulting corpus contains one million sentences and plays a vital role in machine translation research. The data sources that are used for corpus building are Russian/English journalist pages from a number of bilingual web-sites of good quality. Their proposed system for corpus building considers only those pages from web, whose contents are already extracted and properly tokenized. The corpus building procedure in Antonova & Misyurev (2011) consists of tokenization, lemmatization, and detection of potential parallel documents, verification of parallel documents, sentence alignment and filtering out machine translation. The research work in Ijaz & Hussain (2007) discusses various phases in Urdu lexicon development from corpus. For corpus construction, they have collected data from a range of different domains. The most commonly used domains from which Urdu data are retrieved include: sport news, national and international news, data from finance related pages, showbiz, sell and purchase data. For each domain, only one million tokens are included in the corpus. Most data is extracted for two most popular news web site e.g. BBU Urdu and the daily Jang Pakistan. The second source of data, which was considered for Urdu corpus construction are online books and magazines related to required domains. The extracted data was in different encoding systems e.g. the news data were in HTML format while that of the books and magazines were in Inpage format. To maintain integrity of data, the two different formats are converted into standard character encoding scheme i.e. Unicode text files (UTF-16). After conversion, the text is tokenized on the basis of characters like white space, punctuation marks, special symbols etc. Diacritics from the text are removed and at final step, word frequencies are also updated.

Table 2. Analysis of different ML techniques used on different dataset for major NLP tasks

S. No	Name of Dataset	Task	Model	Accuracy	Author(s)
1	ORCHID Corpus	Word Segmentation	CRF	95.79%	Haruechaiyasak *et al.* (2008)
2	ANERcorp	NER	ME		Benajiba *et al.* (2007)
3	Penn-TreeBank	POS	ME	81.57% (Average)	
4	UPenn Chinese Treebank	Word Segmentation	CRF	95%	Peng *et al.* (2004)
5	News corpus (www.jang.com.pk)	POS	SVM	94.15%	Sajjad & Schmid (2009)
6	Wall Street Journal(WSJ) and Brown Corpus	POS	SVM And Naïve Bays	Not mentioned	Gillick (2009)
7	IRL Japanese	NER	SVM	90.03%	Isozaki & Kazawa (2002)
8	Penn-TreeBank	POS	HMM	99.83%	Youzhi (2009)
9	Corpus of Holly Quran	POS and MA	HMM	96%	Elhadj (2009)
10	MUC-7	NER	HMM	90.93%	Todorovic *et al.* (2008)

7. Future direction

This section elaborates future directions in the field of NLP and its application areas. To resolve ambiguity in linguistic knowledge in a better way, we suggest few directions that may help to improve the performance.

First: From literature we have found that natural language processing from the computational perspective has not been as widely investigated with neural network, Naïve Bayes, genetic algorithm (GA), dynamic programming (DP) and evolutionary algorithms. Therefore, in future these mentioned techniques can be studied for building computational models that are able to analyze natural languages for performing useful tasks, e.g; firstly to enable humans and machines to communicate with each other in effective way; secondly to standardize communication system among humans of various dialects; thirdly, to improve existing text or speech processing approaches.

Second: It is also observed from literature that the contribution rate of avant grade ML techniques for major NLP tasks in English as well as for European languages is much higher, as compared to South East Asian languages. Therefore, in future the research community of NLP can increase the contribution of ML techniques for South East Asian languages too.

Third: Cooperative systems or social networks such as Twitter, Face book and Flicker are ubiquitous now a days and plays imperative role in our social life. The users of these systems belong to every walk of life. It is observed that the users of these systems often does not follow syntactical and grammarian rule of the natural languages, while dropping text or expressing their views on particular contents of these systems. The dropped text on these systems is thus full of syntactic and semantic ambiguity. Information extraction from these systems with conventional approaches is more challenging due to high ambiguous text. So in future, more robust and more efficient ML techniques are required for valuable information extraction. Therefore, in future, researchers can investigate the use of ML techniques in social network language processing domain.

Fourth: Performance of ML based approaches largely depend on presence of large amount of training data. Generally performance of ML approaches increases by increasing size of training data, while degrades by decreasing size of training data. Free availability of large corpuses for most of growing languages is a hard issue. Thus unavailability of large annotated corpus leads the researcher to two alternative learning methods: semi-supervised learning and unsupervised learning. So a large annotated corpora once available, natural language processing work with supervised machine learning techniques can move forward.

Fifth: Lot of research work has been conducted for creating reliable and standard linguistic resources for Western languages. In Europe, a remarkable job has been done by the human language technology (HLT) society to standardize the available linguistic resources of European languages. Compared with Western Languages Asian languages are less investigated in terms of linguistic resources. Therefore, in future language resources and tools development for Asian languages using machine learning techniques for low resource languages of sub-continent needs more attention.

Sixth: Although the NLP research community has scalable and reliable rule-based, statistical and hybrid techniques that resolve problems more efficiently in various NLP tasks (e.g. POS, NER, sentence boundary detection, word sense disambiguation, segmentation and WordNet development etc). in future, many other sources of information can be exploited (e.g. Joint learning, transfer learning knowledge bases, unlabeled data, real-world facts). Regarding neural networks' competences in multi output learning, investigating its capabilities in joint learning NLP task is definitely an interesting issue.

8. Conclusion

ML techniques for treating the problem of linguistic knowledge disambiguation have developed remarkably in the last decade. The use of ML techniques for linguistic knowledge disambiguation in the research domain of NPL is a hot research area, gaining attention of NLP research community at a rapid pace. In this paper, we explored various methods that are applied to solve various NLP problems. We reviewed studies on applying different supervised machine learning models to major NLP tasks, which consist of HMM, CRF, maximum entropy (MaxEnt), SVM, Naïve Bays and deep learning. As far as we know, the present work is the first one, which brings discussion to a single search space about ambiguity and its various categories in terms of NLP, various linguistic knowledge concepts, major NLP tasks, machine learning techniques and their corresponding categories. A comprehensive review of different avant grade machine learning models, which are used in literature to address various NLP tasks and finally a brief description of most commonly used dataset and a list of online available corpus are available. This survey paper gives a generic structure and guidelines for developing new ML techniques and methods for address major NLP tasks. This survey paper also provides a clear picture about which model is more often used, as compared to other ones to address major NLP tasks. This study intends to cover all supervised machine learning techniques in context of NLP tasks about different linguistic analysis.

Acknowledgement

The work is supported by Higher Education Commission (HEC), Islamabad, Pakistan.

References:

Abdel Rahman, S., Elarnaoty, M., Magdy, M. & Fahmy, A. (2010). Integrated machine learning techniques for Arabic named entity recognition. International Journal of Computer Science Issues (IJCSI), **7**(4):27-36.

Agarwal, N., Ford, K.H. & Shneider, M. (2005). Sentence boundary detection using a maxEnt classifier. Proceedings of MISC, pp. 1-6.

Ahmed, F. & Nürnberger, A. (2009). Corpora based approach for Arabic/English word translation disambiguation. Speech and Language Technology, **11**:195-214.

Akita, Y., Saikou, M., Nanjo, H. & Kawahara, T. (2006). Sentence boundary detection of spontaneous Japanese using statistical language model and support vector machines. Paper presented at the Interspeech, pp. 1033-1036.

Ammar, W., Dyer, C. & Smith, N.A. (2014). Conditional random field auto encoders for unsupervised structured prediction. Proceedings of the Advances in Neural Information Processing Systems 26(NIPS-2014), pp. 1-9.

Antonova, A. & Misyurev, A. (2011). Building a Web-based parallel corpus and filtering out machine-translated text. Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web. Association for Computational Linguistics, pp. 136–144.

Antony, P., Mohan, S.P. & Soman, K. (2010). SVM based part of speech tagger for Malayalam. Proceedings of IEEE International Conference on Recent Trends in Information, Telecommunication and Computing (ITC), pp. 339-341.

Anwar, W., Wang, X., Li, L. & Wang, X.L. (2007). A statistical based part of speech tagger for Urdu language. Proceedings of International Conference on Machine Learning and Cybernetics, pp. 3418-3424.

Barakat, H., Nigm, E. & Khaled, O. (2014). Statistical modeling of extremes under linear and power normalizations with applications to air pollution. Kuwait Journal of Science, **41**(1):1-19.

Benajiba, Y. & Rosso, P. (2008). Arabic named entity recognition using conditional random fields. In proceedings of Workshop on HLT & NLP within the Arab World, (LREC), pp. 1-7.

Benajiba, Y., Rosso, P. & Benedíruiz, J.M. (2007). Anersys: An arabic named entity recognition system based on maximum entropy. Proceedings of 8th International Conference on Computational Linguistics and Intelligent Text Processing, pp. 143-153.

Borthwick, A. (1999). A maximum entropy approach to named entity recognition. A dissertation in partial fulfillment of the requirement for the degree of Doctor of Philosophy, New York University, pp. 1-115.

Bygate, M., Swain, M. &Skehan, P. (2013). Researching pedagogic tasks: Second language learning, teaching, and testing. Publisher: Routledge, UK.

Danker, F.W. (2000). A Greek-English lexicon of the New Testament and other early Christian literature. Publisher: University of Chicago Press, Chicago, USA.

Daud, A., Khan, W. & Che, D. (2016). Urdu language processing: a survey. Artificial Intelligence Review, 1-33. DOI 10.1007/s10462-016-9482-x.

Deng, L., & Yu, D. (2014). Deep learning. Signal Processing, **7**:3-4.

Ekbal, A. & Bandyopadhyay, S. (2010). Named entity recognition using appropriate unlabeled data, post-processing and voting. Informatica, **34**(1):55-76.

Ekbal, A., & Bandyopadhyay, S. (2009). Named entity recognition in Bengali: A multi-engine approach. Proceeding of the Northern European Journal of Language Technology, pp. 26–58.

Ekbal, A. & Bandyopadhyay, S. (2008). Part of speech tagging in bengali using support vector machine. Proceedings of International Conference on the Information Technology (ICIT, 2008), pp. 106-111.

Ekbal, A., Haque, R. & Bandyopadhyay, S. (2008). Maximum entropy based bengali part of speech tagging. Advances in Natural Language Processing and Applications Research in Computing Science RCS Journal, **33**:67-78.

Ekbal, A., Haque, R., Das, A., Poka, V., & Bandyopadhyay, S. (2008). Language independent named entity recognition in Indian languages. Proceeding of International Joint Conference on Natural Language Processing (IJCNLP), pp. 1-7.

Ekbal, A., Naskar, S.K. & Bandyopadhyay, S. (2007). Named entity recognition and transliteration in Bengali. Lingvisticae Investigationes, **30**(1):95-114.

Elhadj, Y.O.M. (2009). Statistical p-of-speech tagger for traditional Arabic texts. Journal of Computer Science, **5**(11):794-800.

Gillick, D. (2009). Sentence boundary detection and the problem with the US. In Proceedings of Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 241-244.

Gouda, A.M. & Rashwan, M. (2004). Segmentation of connected Arabic characters using hidden Markov models. Proceedings of the IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA.), pp. 115-119.

Han, J., Kamber, M. & Pei, J. (2006). Data mining: concepts and Techniques. Publisher: Elsevier, Amsterdam, Netherlands.

Haruechaiyasak, C., Kongyoung, S. & Dailey, M. (2008). A comparative study on thai word segmentation approaches. Proceedings of 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology,. (ECTI-CON), pp. 1- 4.

Ijaz, M. & Hussain, S. (2007). Corpus based Urdu lexicon development. The proceedings of Conference on Language Technology (CLT07), University of Peshawar, Pakistan, pp. 1-12.

Isozaki, H. & Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. Proceedings of the 19th International Conference on Computational Linguistics (ACL), pp. 1-7.

Jurafsky, D. & James, H. (2000). Speech and language processing an introduction to natural language processing, computational linguistics, and speech. Publisher: Prentice Hall, United States of America.

Kolár, J. & Liu, Y. (2010). Automatic sentence boundary detection in conversational speech: A cross-lingual evaluation on English and Czech. Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 5258-5261.

Lafferty, J., McCallum, A. & Pereira, F.C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the Eighteenth International Conference on Machine Learning, (ICML), pp. 282-289.

Li, Y., Miao, C., Bontcheva, K. & Cunningham, H. (2005). Perceptron learning for Chinese word segmentation. Proceedings of Fourth Sighan Workshop on Chinese Language Processing (Sighan-05), pp. 154–157.

Liu, X., Wei, F., Zhang, S. & Zhou, M. (2013). Named entity recognition for tweets. ACM Transactions on Intelligent Systems and Technology (TIST), **4**(1):1524-1534.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M. & Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. IEEE Transactions on Audio, Speech, and Language Processing, **14**(5):1526-1540.

Low, J.K., Ng, H.T. & Guo, W. (2005). A maximum entropy approach to Chinese word segmentation. Proceedings of the Fourth Sighan Workshop on Chinese Language Processing, pp. 1-4

Luo, X. (2003). A maximum entropy Chinese character-based parser. Proceedings of the 2003 Conference on Empirical Methods in Natural Language, pp. 1-7.

Mohammed, N.F. & Omar, N. (2012). Arabic named entity recognition using artificial neural network. Journal of Computer Science, **8**(8):1285-1293.

Mohit, B., & Hwa, R. (2005). Syntax-based semi-supervised named entity tagging. Proceedings of the Association for Computational Linguistics (ACL 2005) on Interactive Poster and Demonstration Sessions, pp. 57-60.

Morwal, S. & Chopra, D. (2013). NERHMM: A tool for named entity recognition based on hidden Markov model. International Journal on Natural Language Computing (IJNLC), **2**:43-49.

Morwal, S. & Jahan, N. (2013). Named entity recognition using hidden Markov model (HMM): An Experimental Result on Hindi, Urdu and Marathi Languages. International Journal of Advanced Research in Computer Science and Software Engineering, **3**(4):671-675.

Moses, D. (2015). A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ECG data. Kuwait Journal of Science, **42**(2):206-235.

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

Nguyen, C.T., Nguyen, T.K., Phan, X.H., Nguyen, L.M. & Ha, Q.T. (2006). Vietnamese word segmentation with CRFs and SVMs: An investigation. Proceedings of 20th Pacific Asia Conference on Language, Information and Computation (PACLIC 2006), pp. 1-8.

Pandian, S.L. & Geetha, T. (2009). CRF models for tamil part of speech tagging and chunking. Proceedings of International Conference on Computer Processing of Oriental Languages, pp. 11-22.

Patel, C. & Gali, K. (2008). Part-of-speech tagging for Gujarati using conditional random fields. Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pp. 117–122.

Peng, F., Feng, F. & McCallum, A. (2004). Chinese segmentation and new word detection using conditional random fields Proceedings of the 20th International Conference on Computational Linguistics, pp. 1-8.

Qi, Y., Das, S.G., Collobert, R. & Weston, J. (2014). Deep learning for character-based information extraction. Proceedings of European Conference on Information Retrieval, pp. 668–674.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 133-142.

Rehman, Z. & Anwar, W. (2012). A hybrid approach for Urdu sentence boundary disambiguation. International Arab Journal of Information Technology (IAJIT), **9**(3):250-255.

Reynar, J.C. & Ratnaparkhi, A. (1997). A maximum entropy approach to identifying sentence boundaries. Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 16-19.

Saha, S.K., Sarkar, S. & Mitra, P. (2008). A hybrid feature set based maximum entropy Hindi named entity recognition. Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pp. 343-349.

Sajjad, H. & Schmid, H. (2009). Tagging Urdu text with parts of speech: A tagger comparison. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 692-700.

Santos, C.D. & Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp.1818–1826.

Singh, U., Goyal, V. & Lehal, G.S. (2012). Named entity recognition system for Urdu. Proceedings of COLING 2012: Technical Papers, pp. 2507–2518.

Sunny, S., David Peter, S. & Jacob, K.P. (2013). Combined feature extraction techniques and Naive Bayes classifier for speech recognition. Computer Science & Information Technology (CS & IT), pp. 155–163.

Talasiewicz, M. (2009). Philosophy of syntax: foundational topics (Book) 1st ed. Vol. 29. Springer Science & Business Media.

Todorovic, B.T., Rancic, S.R., Markovic, I.M., Mulalic, E.H. & Ilic, V.M. (2008). Named entity recognition and classification using context hidden Markov model. Proceeding of 9th Symposium on Neural Network Applications in Electrical Engineering, (NEUREL 2008), pp. 43-46.

Tomanek, K., Wermter, J. & Hahn, U. (2007). Sentence and token splitting based on conditional random fields. Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pp. 1-9.

Wang, H. & Huang, Y. (2003). Bondec–A Sentence Boundary Detector.CS224N Project, Stanford, CA, USA.

Wenchao, M., Lianchen, L. & Anyan, C. (2010). A comparative study on Chinese word segmentation using statistical models. Proceedings of IEEE International Conference on Software Engineering and Service Sciences (ICSESS), pp. 482 – 486.

Xue, N. (2003). Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing, **8**(1): 29-48.

Yao, L., Sun, C., Li, S., Wang, X. & Wang, X. (2009). CRF-based active learning for Chinese named entity recognition. Proceedings of the 2009 IEEE International Conference on Systems, Man, and

Cybernetics, pp. 1557-1561.

Youzhi, Z. (2009). Research and implementation of part-of-speech tagging based on hidden Markov model. Proceedings of Asia-Pacific Conference on Computational Intelligence and Industrial Applications (PACIIA), pp. 26-29.

Zhang, L., Pan, Y. & Zhang, T. (2004). Focused named entity recognition using machine learning. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1-8.

Zhang, Y., Xu, Z. & Zhang, T. (2008). Fusion of multiple features for chinese named entity recognition based on CRF model Information Retrieval Technology: Springer, pp. 95-106.

Zheng, G. & Tian, Y. (2010). Chinese web text classification system model based on Naive Bayes. Proceedings of the International Conference on E-Product E-Service and E-Entertainment (ICEEE), pp. 1-4.

Zheng, X., Chen, H. & Xu, T. (2013). Deep learning for Chinese word segmentation and pos tagging. Proceedings of the 2013 Conference on Empirical Methods in Natural Language (EMNLP-ACL-2013), pp. 647–657.

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

Submitted **:** 29/01/2015
Revised **:** 03/06/2015
Accepted **:** 06/06/2015

دراسة إستقصائية لأحدث نماذج تعلم الآلة في سياق معالجة اللغات الطبيعية

[1,*] وهاب خان، [1,2]علي داود، [1]جمال ناصر، [1]تيهمينا أمجد

[1]قسم علوم الحاسب وهندسة البرمجيات، IIU، إسلام آباد 44000، باكستان
[2]كلية الحوسبة وتقنية المعلومات، جامعة الملك عبدالعزيز، جدة، المملكة العربية السعودية
[*]المؤلف: wahab.phdcs72@iiu.edu.pk

خلاصة

تعلم الآلة والتقنيات الإحصائية تعد من أدوات التحليل القوية التي ما زالت لم يتم إدراجها في الحقل الجديد المتنوع الأنظمة المعروف بإسم معالجة اللغة الطبيعية (NLP) أو اللغويات الحاسوبية. المعرفة اللغوية قد تكون غامضة أو تحتوي على غموض. وبالتالي، يتم تنفيذ مختلف المهام في معالجة اللغات الطبيعية من أجل حل الغموض في الكلام واللغة.

وتعتمد التقنيات الحالية للتعلم الخاضع للإشراف في معالجة اللغات الطبيعية على نماذج ماركوف(Markov) المخفية ، والحقل العشوائي المشروط (CRF)، و نماذج الحد الأقصى للإنتروبيا (MaxEnt)، مصفوفة الدعم الآلي (SVM)، (Naive Bays)، والتعلم العميق (DL) .

والهدف من هذه الورقة هو تسليط الضوء على الغموض في معالجة النطق واللغة، وتقديم نظرة عامة موجزة عن الفئات الأساسية من المعرفة اللغوية، و مناقشة نماذج مختلفة لتعلم الآلة وتصنيفها إلى فئات مختلفة، وأخيرا لتوفير مراجعة شاملة لأحدث نماذج تعلم الألة بهدف أن ينظر الباحثون في هذه التقنيات واعتمادا عليها يتم تطوير تقنيات جديدة متقدمة. في هذه الدراسة استعرضنا أيضا كيف يمكن استخدام نماذج تعلم avantgrade machine في هذه المعضلة.

الكلمات المفتاحية: الغموض؛ المعرفة اللغوية؛ تعلم الآلة؛ معالجة اللغات الطبيعية؛ التعليم تحت إشراف.