Urdu Word Segmentation using Machine Learning Approaches

Sadiq Nawaz Khan[1], Khairullah Khan[2]
Department of Computer Science
University of Science & Technology Bannu, Bannu,
Pakistan

Asfandyar Khan[4]
Institute of Business and Management Sciences
University of Agriculture
Peshawar, Pakistan

Wahab Khan[3]
Department of Computer Science & Software Engineering
International Islamic University
Islamabad, Pakistan

Fazali Subhan[5]
Department of Computer Science
National University of Modern Languages
Islamabad, Pakistan

Aman Ullah Khan[6], Burhan Ullah[7]
Department of Computer Science
University of Science & Technology Bannu
Bannu, Pakistan

***Abstract*—Word Segmentation is considered a basic NLP task and in diverse NLP areas, it plays a significant role. The main areas which can be benefited from Word segmentation are IR, POS, NER, sentiment analysis, etc. Urdu Word Segmentation is a challenging task. There can be a number of reasons but Space Insertion Problem and Space Omission Problems are the major ones. Compared to Urdu, the tools and resources developed for word segmentation of English and English like other western languages have record-setting performance. Some languages provide a clear indication for words just like English which having space or capitalization of the first character in a word. But there are many languages which do not have proper delimitation in between words e.g. Thai, Lao, Urdu, etc. The objective of this research work is to present a machine learning based approach for Urdu word segmentation. We adopted the use of conditional random fields (CRF) to achieve the subject task. Some other challenges faced in Urdu text are compound words and reduplicated words. In this paper, we tried to overcome such challenges in Urdu text by machine learning methodology.**

Keywords*—Part-of-speech (POS); NER; word segmentation; information retrieval; Natural Language Processing (NLP); conditional random fields (CRF)

I. Introduction

Natural Language Processing (NLP) is a key area for research in almost every language of the world. In NLP computers are trained in such a way that can easily understand and manipulate human language text or speech. NLP researchers are trying to produce such a knowledge that how human beings understand and use natural language. They use applicable tools and procedures that can be technologically advanced to make computer systems cognize and operate natural languages and achieve the desired tasks. NLP fundamentals lie in various disciplines such as information and computer sciences, electronic and electrical engineering, linguistics, artificial intelligence (AI), mathematics and psychology, etc. [1]. NLP applications consist of various fields of studies, such as text processing and summarization, user interfaces, CLIR (cross-language information retrieval), speech recognition, AI and word segmentation etc. Recognition of valuable and relevant documents from a large collection with respect to the desired query is information retrieval (IR). The technique which is used to process document or collection of documents for identification of events or entities which have been pre-specified is information extraction (IE). Information extraction (IE) is a technique which processes a document, or collection of documents, to identify pre-specified entities or events.

Word Segmentation has significant role in all NLP applications. It has the ability of dividing and separation of written text into meaningful units which are usually known as words. Words boundaries in a spoken language can be identified by word segmentation. Hindi like languages attracted researcher's attention during recent years. Especially on web Urdu language is going to become a key part of Asian languages [2]. Informational retrieval (IR) and Data Mining (DM) need a detailed knowledge of NLP with responsibilities of the relationship exploration, topic categorization, event extraction and sentiment analysis, etc. NLP significance such as part-of-speech (POS) tagging, morphological analysis, named entity recognition, stop words removal, parsing and shallow parsing have significant importance in all NLP systems [3]. Urdu word segmentation problem is not unadorned as some of the other Asian languages, in which space is used for word demarcation, but it has not consistently been used. The use of space gives rise to both space omission and space insertion problems in Urdu text [4] and [5]. The Space

omission problem e.g. the Urdu word "انکا" which is actually a combination of two words but the system treats it as a single word. Such Segmentation in Urdu text is handled with the application of Urdu-Devnagri transliteration system [6]. The Space Insertion problem e.g. the word "عقل مند"(Aqalmand, Intelligent) is actually one word but when segment it will be treated as two words i.e. عقلand مند which is handled by a two-stage system [7]. Hindi-Urdu transliteration issues are briefly discussed by [8] and [9]. Simple, compound and complex words are segmented for Sindi language using three layers [10]. A complete survey of techniques regarding Urdu-Arabic Word Segmentation and also their challenges have discussed by [11].

II. Literature Review

Nowadays different languages use different techniques for word segmentation problem so far. These techniques are used by NLP researchers and have deduced better results from each one. The existing techniques for word segmentation in NLP are Dictionary/rule-based, statistical/machine learning and hybrid approaches.

A. Existing Techniques

There are some techniques which are commonly used for word segmentation and some are not widely used yet. The detail of these techniques is given below:

1) Rule-Based Techniques: Rule-based techniques are set of rules or pattern which are used to perform various NLP tasks. Rule-based approaches were constructed manually by linguistics experts. This approach was used by [12] for chines word segmentation. They also show a transformation-based algorithm for improving the output of the system. As Urdu, Chines, Japanese and Thai etc have not delimited by spaces, therefore word segmentation is how much difficult as compared to other western languages like English etc. Word segmentation for Thai language using rule-based technique was presented by [13]. An Urdu stemmer namely "Assas Band" developed by [14] is based on rule-based. Assas-Band firstly removes the prefix from the stem and then postfix and finally stem is extracted with the accuracy of 91.2%. Urdu online handwriting recognition system provided by [15]. Author in [16] has used the rule-based technique for Name Entity Recognition in Urdu. Urdu word segmentation using this approach is done by [5].

2) Machine Learning/Statistical Techniques: Machine learning approach is much better than rule-based approaches although this technique is not commonly used for word segmentation. These techniques use learning algorithms which are capable of defining a function that takes input samples to a range of output values. A corpus is constructed for these approaches in which word boundaries are explicitly defined. Statistical models are formed containing features of the words which have been surrounded by boundaries. Supervised statistical learning is one of the most current dominant technique in NLP. This approach automatically induces rules from training data. Machine learning algorithms consist of intelligent modules. Different machine learning models have been discussed by [17]. In order to carry out major NLP task using statistical approaches, it incorporates stochastic and probabilistic methods. A two-stage word segmentation system for handling space insertion problem in Urdu by [7] is done using the statistical-based technique. The space omission problem in Urdu word segmentation using this approach has been used by [6].

3) Hybrid Approaches: Hybrid techniques are the combination of features of rule-based and statistical techniques. Authors in [18] presented a hybrid approach for Urdu sentence boundary disambiguation comprising of unigram statistical model and rule-based algorithm. These results better than rule-based and statistical based approaches. Hybrid technique for segmentation presented by [19] uses top-down mechanism for line segmentation and bottom-up design for segmenting the line into ligatures. The accuracy result was achieved 99.2% using this technique.

III. Urdu Language

Urdu is the National language of Pakistan. The hand-held devices such as mobiles phones, etc. have been successfully using everywhere but the software they provide for user input is mostly in English and in Pakistan, it is difficult for a common man to communicate in English easily. In order to facilitate Urdu speakers and writer and reduce the difference between the common man and the new technology, Urdu NLP systems are required. We have tried to bridge this gap by using machine learning approach for segmentation of Urdu text.

A. Urdu Writing Style

Urdu is not scripting language although it is written in cursive Arabic script. Arabic script has many traditional writing styles such as Naskh (mostly used for the Arabic language), Taleeq, Kufi, Divani, Sulus, Riqa, etc. As Nastalique is complex writing style but it is novel and robust and most commonly used for Urdu writing. Nastaleeq is character based, bidirectional (mainly right to left), diagonal, on-monotonic, cursive, context-sensitive writing system with a significant number of marks (dots and other diacritics).

1) Urdu Characters: Urdu has 50 consonants in which 35 are simple and 15 are aspirated. There are 15 diacritical marks and 1 character for nasal sound. Consonant letters, vowels, diacritic marks, numerals, punctuations and few superscripts signs support Urdu text. Urdu text can be written with simple characters or characters with diacritical marks. Both format conveying same meaning but the only difference is in writing and oral saying e.g. the Urdu word having simple character "ٹولکفل" is same in oral saying as "ٹُولکِفل" which have two diacritic marks i.e ُ and ِ. For segmentation, such diacritic marks will have to remove first. Table I shows the Urdu digits and characters, while Table II shows some other characters which are not counted as part of the alphabet, punctuation marks, signs, and symbols of Urdu text.

TABLE I. URDU DIGITS AND ALPHABETS

Urdu Writing Style	Digits & Alphabets		
	Numbers & Characters	Numbers	Characters
		۹ ۸ ۷ ٦ ٥ ٤ ٣ ٢ ١ ٠	ا ب بھ پ پھ ت تھ ٹ ٹھ ث ج جھ چ چھ ح خ د دھ ڈ ڈھ ر ڑ ڑھ ز ژ س ش ص ض ط ظ ع غ ف ق ک کھ گ گھ ل م ن و ہ ی ے

TABLE II. TABLE TYPE STYLES

Urdu Writing Style	Diacritics, punctuation mark, signs & symbols		
	Characters not counted as part of alphabet/diacritics	Punctuation marks	Signs & Symbols
	◌ٌ ◌ً ھ ں آ ۂ ۓ ئ ؤ ء ◌َ ◌ّ ◌ِ ◌ُ ◌ٖ ◌ٗ ◌ٰ ◌ٍ ◌ْ ◌ٓ	، ؛ ؟ ۔ ٫ ،	؏ ؎ ؁ ؀ ؂ ◌ؔ ◌ؓ ◌ؒ ◌ؑ ◌ؐ ﷽

2) *Joiners*: Urdu script is cursive and characters are joined with neighbor within a word and acquire different shapes. Such characters are known as joiners. Joiners have four-way shaping i.e. initial, medial, final and isolated form. Table III shows some examples of four-way shaping form of joiners and Table IV shows joiner characters of Urdu text.

TABLE III. FOUR-WAY SHAPING OF JOINERS

Urdu Writing Style	Four-way Shaping of Urdu Joiners			
	Final	Medial	Initial	Isolated
	بب	ببب	با	ب
	بت	بتب	تا	ت
	بٹ	بٹب	ٹا	ٹ
	بی	بیب	یا	ی

TABLE IV. JOINERS IN URDU

Urdu Writing Style	Joiners in Urdu
	Joiners
	ب ت ۃ ث ٹ ج چ ح خ ہ س ش ص ض ط ظ ع غ ف ق ک گ ل م ن ء ی

3) *Non-Joiners*: Some Urdu characters are not joined with the neighbor ones, such characters are referred to as non-joiners. Non-joiners have only two forms i.e final and isolated. The following Table V shows some examples of the final and isolated forms of non-joiners whereas the Table VI shows non-joiner characters of Urdu text.

TABLE V. FORMS OF NON-JOINERS

Urdu Writing Style	Forms of Non-Joiners		
	Urdu non-joiners	Final	Isolated
		با	ا
		بد	د
		ظر	ر
		کو	و
		کے	ے

TABLE VI. NON-JOINERS IN URDU

Urdu Writing Style	Non-Joiners in Urdu
	Non-Joiners
	آا د ڈ ذ ر ز ڑ ژ و ے

B. Urdu Linguistics Resources

Urdu lexical resources are a necessary part of every NLP system for the computational processing of Urdu language. In Pakistan the area of applied linguistic such as English language teaching (ELT) and sociolinguistic are the two highly focused fields by the researchers. Very trivial study has been reported in respect of descriptive and theory-based linguistics and there has an evenly finite capability in that area in Pakistan. For the purpose stated, one of the leading "Essential Urdu Linguistic Resources project" is concentrating on building up indispensable Urdu lingual resources and tools by ramping up research capability in grammatic and semantic studies. This coaction will assist research community to abloom the area of linguistics inside Pakistan.

The Urdu corpus and lexical resources developed for Urdu has been discussed by [20] are listed below:

1) *Urdu Encoding Scheme*: The computer keyboard is used as input device for entering data to the computer. It contains characters, numbers, functional keys and symbols etc. Special encoding technique is used when the computer gets input. Character encoding is the process of assigning a unique number to each character of the language. This code is generated internally in a computer system. For Urdu language, different encoding schemes have been developed but for standardization of encoding scheme, no effort was undertaken.

2) *UTF*: The Unicode organization is responsible to develop and assign a unique character encoding scheme for digital text of almost all languages of the globe. The most general Unicode character scheme that are commonly in use are UTF-8, UTF-16, and UTF-32. Majority operating systems are based on UTF-16 encoding scheme. This encoding scheme is adopted as worldwide encoding scheme and is capable to map all known characters.

3) *Urdu Zabta Tahti (UZT)*: As there are no industry standards for coding in Urdu, similar to ASCII standard for English, therefore, it needs much attention. For this purpose Urdu Zabta Tahti (UZT) version 1.01 by [21] is a standard code page for Urdu. The Government of Pakistan has accepted UZT version 1.01 as a standard code for Urdu.

4) *Urdu Text Corpus*: In 2002 Becker and Riaz released the first publicly freely available Urdu dataset to promote research activities in Urdu. In its development the contents of 7000 news articles were used which was extracted from BBC

Urdu URL. The Becker and Riaz dataset contains very reach contents and is considered feasible for majority of ULP tasks such as Part of speech tagging, named entity recognition and so on.

EMILLE project has made Urdu corpus for the first time by [22]. The corpus has 200,000 words of English text translated into Urdu etc. and 512000 words of spoken Urdu and 1640000 words of Urdu text.

5) CLT Conference: In Pakistan the Society for Natural Language Processing (SNLP) has taken initiative steps to arrange a series of international conference, namely, Conference on Language and Technology (CLT) with the objectives to abide students, researchers of various universities and research institutions to share research ideas and to promote research culture in Pakistani and South Asian languages.

6) SNLP: Recently researcher has shown growing interest in the computational processing of Urdu digital text in Pakistan. In Pakistan there are assorted number of organizations and individuals which perform research activities in isolated manners and there exists no coordination among various organizations and individuals.

An integrated exertion is necessary to bring in them in collaborative platform to present ideas and pass around information. SNLP renders a research platform for organizations and individual researchers for this aim.

These days more than 60 languages are mouthed in Pakistan; hence we can state that Pakistan symbolizes a diverse still adhesive lingual and cultural environment. Lot languages are interconnected and several are generally mouthed crosswise territorial bounds. Hence, there is a demand to build up a basic platform to draw together the research community processing these languages.

IV. Conditional Random Fields

CRF is a machine learning algorithm, which is widely used in Natural Language Processing (NLP) tasks e.g. word segmentation, sequential labeling, Name Entity Recognition and so on. Conditional Random Fields (CRFs) are undirected graphical models used to calculate the conditional probability of values on designated output nodes given values on designed input nodes. CRF has several advantages over Hidden Markov models and stochastic grammar models (Lafferty, McCallum, & Pereira, 2001) and defines a CRF on X and random variable Y as follows:

Let the graph $G = (V, E)$ such that $Y = (Y_y)$vεv) so that Y is indices by the vertices of G. Then (X, Y) is conditional random field when the random variable Y_v, conditioned on X, obey the Markov property with respect to the graph: $p\,(Y_y/X, Y_{W,}\, w{\sim}v)$ means that w and v are neighbors in G. For sequence tagging tasks, the LDCRF (Latent-dynamic random fields) or DPLVM (Discriminative Probabilistic Latent Variable Models) are a type of CRFs for sequence tagging tasks. These models are known as latent variable modela that are trained discriminatively. According to LDCRF let a given sequence of observations say, $X = x_1, x_2, x_3,\ x_n$ one of the tagging task but here the problem arises that how to assign sequence of labels and this problem should be solved by the model let $Y = y_1, y_2, y_3, \ldots y_n$, be a labels sequence. In ordinary linear-chain CRF, latent variables 'h' is inserted between x and y rather than directly modeling $P(Y/X)$. It uses chain rule probability.

$$\mathrm{P(Y/X)} = \sum_{\mathrm{h}} \mathrm{p(Y/h\,,X)P(h/X)} \qquad (1)$$

Suppose $\mathrm{x_{1:n}}$ is a sequence of Urdu words in a sentence with name entities $\mathrm{z_{1:n}}$. According to linear chain CRF, the conditional probability is as:

$$\mathrm{P\,(z_{y:n}\Big/x_{1:n}) = 1\Big/Z\exp\left(\sum_{n=1}^{N}\sum_{i=1}^{F}\lambda_i\ f_i(\,z_{n-1},\,z_n,\,z_{1:n},n)\right)} \qquad (2)$$

Where the normalization factor Z is calculated as under

$$= \sum_{z_{1:n}} \exp\left(\sum_{n=1}^{N}\sum_{i=1}^{F}\lambda_i\ f_i(\,z_{n-1},\,z_n,\,z_{1:n},n)\right) \qquad (3)$$

V. Name Entity Recognition

NER was first introduced in 1995 as part of MUC-6 (Message Understanding Conference). Later on, in 1996, the MET-1 conference introduced the name entity recognition in the non-English text. Name entity is one of the prior tasks in NLP. Named entity recognition consists of identifying within sentence words or sequences of adjacent words belonging to a certain class of interest or it classifies proper nouns into its predefined categories such as a person, time, date, brand names, quantities, monetary values, percentages, abbreviations, location, organization, etc. For each class of interest, the labeling distinguishes between the first word in the named entity and the following words in the named entity. Words not belonging to any class of interest are labeled as O (other). Name entity recognizer is the software which labels sequence of words in a text. Word segmentation has been applied in several tasks e.g. NER, IR, automatic speech recognition, machine translation, etc. There are two types of approaches to utilize word segmentation in such tasks: pipelining and joint-learning. The pipeline approach creates word segmentation first and then feeds the segmented words into the subsequent task(s). The joint-learning approach trains a model to learn both word segmentation and the subsequent task(s) at the same time. Many NER types of research are based on word segmentation and even Part-Of-Speech (POS) tagging. The relationship between them is described in Fig. 1.

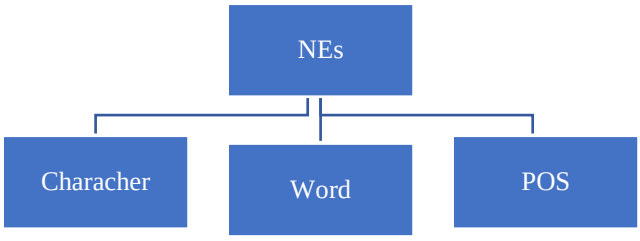

Fig. 1. NER model for segmentation.

The main goal of NER is to recognize the name entities and then resolve the ambiguities from them. Two types of ambiguities are common in names i.e. structural ambiguity and semantic ambiguity has been discussed briefly by [23]. They implemented a module for proper names recognition. Considerable work has been done for NER in western languages such as English, etc. but the interest for NER in South Asian languages has not been developed so far. The main reason is lack of technologies for South Asian languages. Urdu is one of the most important languages of South Asia and a lot of efforts are going on for the development of this language throughout the world especially in Pakistan because Urdu is a national language of Pakistan. The first effort in NER for South Asian languages was made by [24], who highlighted the main challenges facing NER for the Urdu language. They created Becker-Riaz Urdu corpus for the first time as there was no other resource available at that time. In IJCNLP conference 2008, a comprehensive attempt for NER was made for South Asian languages. Many experiments have done for NER in Urdu which uses CRF up to some extent, but need more attention and deep study while using CRF as a module for NER in Urdu.

CRF Classifier provides a general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models for any task. In our work, the NER structured as to consider the following Urdu sentence.

پاناما کیس کے فیصلہ سے قبل گیلپ پاکستان نے عوامی رائےجانی۔

[پاناما کیس: NOR] [کے: NOR] [فیصلہ: NOR] [سے: NOR] [قبل: NOR] [گیلپ: PER] [پاکستان: LOC] [نے: NOR] [عوامی: PER] [رائے: NOR] [جانی: NOR]

VI. Challenges in Urdu Word Segmentation

Urdu word segmentation faces different challenges such as Space omission problem, space insertion problem, compound words, reduplicated words, affixations and English Abbreviations. All these challenges of word segmentation are briefly discussed below.

A. Space Insertion Problem

When space is inserted in between two words of Urdu then the space insertion problem arises. In handwritten Urdu text, there is no space inserted in between words and are briefly discussed by [8], [5] and [6]. When the ending character of the word is joiner then space must be inserted to separate the words otherwise they make a miss-understandable form which the system does not recognize it however the native speaker of Urdu can understand. E.g. consider the Urdu words داخلہ فارم (dahla form, Admission Form) ہم منصب (hum mansab, counterpart) is a combination of four words but semantically these are two words. Now if we remove the spaces in between the words then the above words look like ہممنصب، داخلہفارم which having visually incorrect shape, means in such like cases space must be inserted in between the words otherwise system will not recognize such words. But the problem hereby arises that if we put space in between such words then it is also difficult for a system to take it as a single word because such words are a combination of different words. Similarly, consider the whole sentence, "اقرارکرنایاانکارکرنا" (iqrar karna ya inkar karna, accept or refuse) having five separate words اقرار، کرنا، یا، انکار، کرنا can easily understand and segmented by the native speaker of Urdu. But the system will take this whole sentence as a single word. Space insertion problem causes due to multiple reasons which have been briefly discussed by [5].

B. Space Omission Problem

When space is omitted in such a place where it should be inserted for the appropriate form of the word, then space omission or space exclusion problem arises. Space omission in Urdu text is also a challenging task for word segmentation. If a word ends with a joiner character then it should be separated by a space otherwise it will append to next word which then gives visually incorrect shape. Consider the word شاہی قلعہ (Shahi Qilla), if space is omitted then it will look like شاہیقلعہ having a visually incorrect shape for reader and system as well. But there are some words in which if space is omitted then they do not lose their meaning and have correct shape also. Consider the words: آپ کا(yours) ،(will do) کرے گی کے لئے، (for), (narrate) بیان کریں , اس وقت(at that time), after omitting the space in between these words they make the forms: آپکا، کریگی، کیلئے، بیانکریں، اسوقت all these shapes are acceptable and understandable by the system and the native speakers (Durani & Hussain, 2010). Thus we can say that space is not always used as a word boundary in Urdu. One of the considerable approach for handling space omission problem in Urdu word segmentation is used by (Lehal, 2010), which is based on Urdu-Devnagri transliteration system, in which Urdu words are translated into Hindi Devnagri and then segmented.

C. Compound Words

Compound word is the combination of two or more lexemes to form another lexeme [25]. Compounding is the process in which new units of thought are formed. [8] have categorized the compound words into three categories.

- AB
- A o B
- A e B

The examples of Urdu words in the above formats are جیل خانہ(jail khana, Prison), محنت و عظمت(mehnat o azmat, hardworking and greatness) and حالتِ زار(halat-e-zar, bad condition). In our system, these compound words are handled while doing word segmentation.

D. Reduplicated Words

Reduplicated are those words in which one word/morpheme occurs twice consecutively. Jawaid & Ahmed, 2009 has discussed the Urdu reduplicated words: دن بدن(din ba din, day by day), کبھی کبھی(Kabhi Kabhi, whenever). By observing the above two reduplicated words it is concluded that in reduplication one word is repeated twice or a morpheme is added to that word and make reduplicated word e.g. in دن بدن word the morpheme ب is added to the repeated word دن. The reduplicated words will treat by the system as separate orthographic words (Durrani & Hussain, 2010).

In Urdu word segmentation such words need proper attention and in our work, these words are handled up to some extent.

E. Affixations

In Urdu text affixation (prefixes and suffixes) are used e.g انتھک(anthak, tireless) is an example of prefixes which should be a single word [3]. Similarly, the examples of words with suffixes بد اخلاق(bad akhlaq, bad character), با وقار(ba Waqar, honorable) etc should also consider as single words [14].

F. English Words

Urdu is a language which borrows words from other languages such as Arabic, Farsi, Greek, Latin, and English etc. Abbreviations of English in Urdu writing needs a space/dash character in between the words [8], e.g. Ph.D. (پی ایچ ڈی) or (پی-ایچ-ڈی), M.Phl (ایم فِل) etc.

VII. Urdu Word Segmentation Model

The proposed CRF based Urdu word segmentation model makes use of named entities and POS information of words as a feature for the subject task.

For POS tag information we used CLE POS tagged corpus and for NE information we used the UNER dataset [26]. The UNER dataset contains only NE tags since POS information of particular words provides important information about the basis of the word. Therefore, to make the UNER dataset more informative for feature learning task we first assigned POS tags to each word of the UNER dataset. For this purpose, we make legal use of CLE POS tagged corpus. The assignment of POS task is achieved with help of longest maximum matching technique.

After POS tag assigned to the whole UNER dataset CRF model is trained on this UNER dataset containing both POS and NE tags. This new UNER dataset is used to generate a model file with help of feature set provided in below table. The resultant training model file of CRF is then used along with lexical dictionary file for testing test data. The following Table VII shows the feature template for our proposed model.

TABLE VII. Feature Template for Proposed Model

Features	
Feature Template	***Description***
U01:%x[-1,0]	N-1 token
U02:%x[0,0]	Current token
U03:%x[1,0]	N+1 token
U04:%x[-1,0]/%x[0,0]	N-1 word and N+1 token
U05:%x[0,0]/%x[1,0]	Current token and N+1 token
U06:%x[-1,0]/%x[1,0]	N-1 token and N+1 token
U07:%x[-1,1]	POS tag of N1 token
U08:%x[0,1]	POS tag of the current token
U09:%x[1,1]	POS tag of N+1 token

Fig. 2 below shows the graphical depiction of proposed CRF.

A brief summary of the steps is below:

- UNER dataset is pre-processed
- CLE corpus is pre-processed
- POS tags are assigned to UNER dataset using Longest maximum matching techniques
- The new UNER dataset is then modeled in CRFSharp package requirements
- CRF is trained using the feature template
- The model file is generated
- Test data is tested for word segmentation task against the model files and dictionary files
- Output is generated
- Result is calculated
- Results are averaged

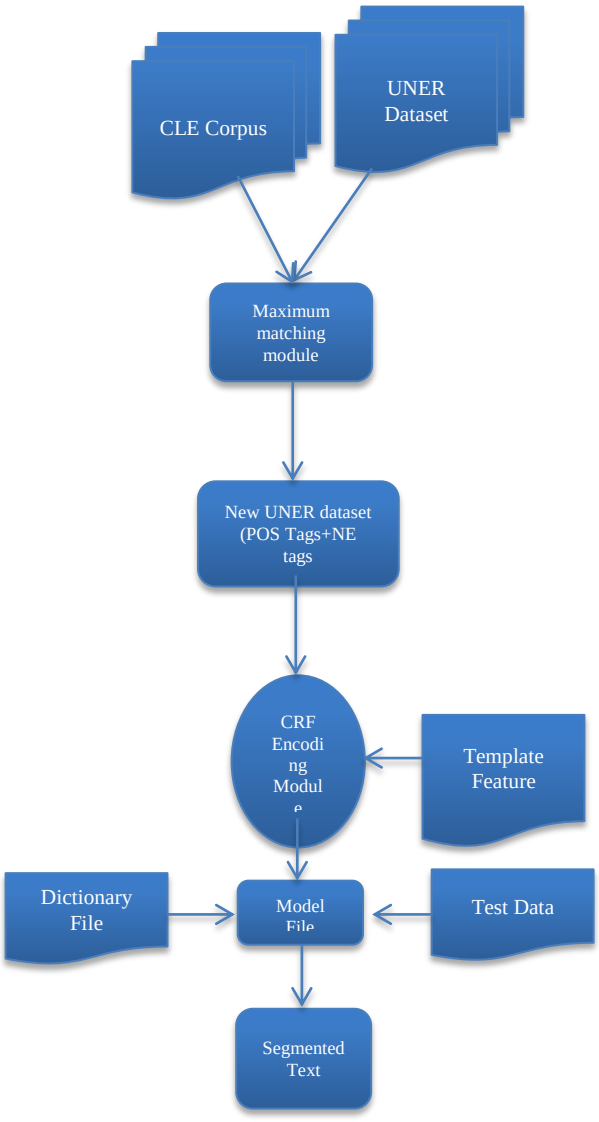

Fig. 2. Graphical depiction of proposed CRF model.

VIII. Experiments and Evaluations

To evaluate the performance of our proposed system we used WordSeg [1] libraries a C# implementation. Training corpus contains 320413 Urdu words, in which compound, reduplicated and foreign words are also included. The overall performance of the system is evaluated using Precision, Recall and F-measure (F-score). Precision and Recall are inversely related to each other as Precision increases, Recall decreases and vice versa. F-measure is the value gained from calculating the harmonic mean of Precision and Recall. For testing Urdu text was taken from [2] BBC site. The text was in the form of sentences in four cases. Table VII shows the Precision, Recall and F-score values for the test data. The tested Urdu text is in the form of sentences and the number of sentences and words for the four cases are given in Table VIII.

TABLE VIII. Tested Text Results

Results	**Precision, Recall & F-score values of Urdu tested text**			
Tested Text		***Precision***	***Recall***	***F-Score***
Sentence	***Words***			
1	23	100%	50%	67.5%
3	50	94%	51%	66%
6	99	94%	51%	66%
31	497	96%	50%	65.7%

The results show the average values of Precision, Recall and F-score for all the tested four cases are 96%, 50.7%, and 66.3%, respectively. It was observed that increasing the training data for Urdu word segmentation improves the results as well. The main challenges in Urdu word segmentation i.e. space insertion and omission problems, reduplication, compound words and foreign words are covered up to some extent depending on the training corpus.

In this study, we considered the research work of [27] as baseline work. The comparison of the proposed system with baseline work is shown in Table IX:

TABLE IX. Comparison of proposed CRF Model with Baseline Approach

Results Comparison	**Comparison of Proposed CRF Model with Baseline Approach**				
Approaches	***Problem Addressed***	***Tested Text***	***Correctly Segmented Words***	***Uncorrected Segmented Words***	***Accuracy***
Baseline Approach	Space Omission	11,995	11,723	272	97.2%
Proposed CRF Approach	Space Omission, Deletion, compound, Reduplicated, Abbreviation, English Words	3,161	3,118	43	98.6%

[1] https://github.com/zhongkaifu/CRFSharp
[2] http://www.bbc.com/urdu/sport

IX. Conclusions

In this paper we have presented a system for solving Urdu word segmentation using machine learning approaches i.e. CRF algorithms. The task of Urdu word segmentation is more challenging as compared to other Asian languages because of space problems in between the words. The objective of this study was to present ML based new system for Urdu word segmentation in which both the main issues of segmentation i.e. space insertion and space deletion as well as compound words and reduplicated words, are handled up to some extent. We believe that the proposed word segmentation system is more advanced system when compared to previous systems as it addresses simultaneously space insertion, space deletion, compound words and reduplicated words challenges.

References

[1] G. G. Chowdhury, "Natural language processing," Annual review of information science and technology, vol. 37, pp. 51-89, 2003.

[2] S. Mukund, R. Srihari, and E. Peterson, "An Information-Extraction System for Urdu---A Resource-Poor Language," ACM Transactions on Asian Language Information Processing (TALIP), vol. 9, p. 15, 2010.

[3] A. Daud, W. Khan, and D. Che, "Urdu language processing: a survey," Artificial Intelligence Review, pp. 1-33, 2016.

[4] N. Durrani, "Typology of word and automatic word Segmentation in Urdu text corpus," 2007.

[5] N. Durrani and S. Hussain, "Urdu word segmentation," in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010, pp. 528-536.

[6] G. S. Lehal, "A word segmentation system for handling space omission problem in urdu script," in 23rd International Conference on Computational Linguistics, 2010, p. 43.

[7] G. S. Lehal, "A two stage word segmentation system for handling space insertion problem in Urdu script," analysis, vol. 6, p. 7, 2009.

[8] B. Jawaid and T. Ahmed, "Hindi to Urdu conversion: beyond simple transliteration," in Conference on Language and Technology, 2009.

[9] A. Malik, L. Besacier, C. Boitet, and P. Bhattacharyya, "A hybrid model for Urdu Hindi transliteration," in Proceedings of 2009 Named Entities Workshop: Shared Task on Transliteration, 2009, pp. 177-185.

[10] J. Mahar, H. Shaikh, and G. Memon, "A Model for Sindhi Text Segmentation into Word Tokens," Sindh University Research Journal-SURJ (Science Series), vol. 44, 2012.

[11] A. Mahmood, "Arabic & Urdu Text Segmentation Challenges & Techniques," vol. IV, pp. 32-34, 2013.

[12] D. D. Palmer, "A trainable rule-based algorithm for word segmentation," in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 1997, pp. 321-328.

[13] Y. El Hadj, I. Al-Sughayeir, and A. Al-Ansari, "Arabic part-of-speech tagging using the sentence structure," in Proceedings of the Second

International Conference on Arabic Language Resources and Tools, Cairo, Egypt, 2009.

[14] Q.-u.-A. Akram, A. Naseer, and S. Hussain, "Assas-Band, an affix-exception-list based Urdu stemmer," in Proceedings of the 7th workshop on Asian language resources, 2009, pp. 40-46.

[15] S. Malik and S. A. Khan, "Urdu online handwriting recognition," in Emerging Technologies, 2005. Proceedings of the IEEE Symposium on, 2005, pp. 27-31.

[16] K. Riaz, "Rule-based named entity recognition in Urdu," in Proceedings of the 2010 named entities workshop, 2010, pp. 126-135.

[17] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," Kuwait Journal of Science, vol. 43, 2016.

[18] Z. Rehman and W. Anwar, "A hybrid approach for urdu sentence boundary disambiguation," Int. Arab J. Inf. Technol., vol. 9, pp. 250-255, 2012.

[19] G. S. Lehal, "Ligature segmentation for Urdu OCR," in 2013 12th International Conference on Document Analysis and Recognition, 2013, pp. 1130-1134.

[20] S. Hussain, "Resources for Urdu Language Processing," in IJCNLP, 2008, pp. 99-100.

[21] S. Hussain and M. Afzal, "Urdu computing standards: Urdu zabta takhti (uzt) 1.01," in Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International, 2001, pp. 223-228.

[22] R. Carter and J. McRae, Language, literature and the learner: Creative classroom practice: Routledge, 2014.

[23] N. Wacholder, Y. Ravin, and M. Choi, "Disambiguation of proper names in text," in Proceedings of the fifth conference on Applied natural language processing, 1997, pp. 202-208.

[24] D. Becker and K. Riaz, "A study in urdu corpus construction," in Proceedings of the 3rd workshop on Asian language resources and international standardization-Volume 12, 2002, pp. 1-5.

[25] R. W. Sproat, Morphology and computation: MIT press, 1992.

[26] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "Named Entity Dataset for Urdu Named Entity Recognition Task," Organization, vol. 48, p. 282, 2016.

[27] R. Rashid and S. Latif, "A dictionary based urdu word segmentation using maximum matching algorithm for space omission problem," in Asian Language Processing (IALP), 2012 International Conference on, 2012, pp. 101-104.