Urdu Word Segmentation using Machine Learning Approaches

Sadiq Nawaz Khan¹, Khairullah Khan²
Department of Computer Science
University of Science & Technology Bannu, Bannu, Pakistan

Asfandyar Khan⁴
Institute of Business and Management Sciences
University of Agriculture
Peshawar, Pakistan

Wahab Khan³
Department of Computer Science & Software Engineering
International Islamic University
Islamabad, Pakistan

Fazali Subhan⁵
Department of Computer Science
National University of Modern Languages
Islamabad, Pakistan

Aman Ullah Khan⁶, Burhan Ullah⁷
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 "لک" 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 "لاقمند"(Aqalmand, Intelligent) is actually one word but when segment it will be treated as two words i.e. لاقم excursion 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 Nastaleeq 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, punctuation 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 “ذ" is same in oral saying as “ذ" 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.
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**

<table>
<thead>
<tr>
<th>Urdu Writing Style</th>
<th>Final</th>
<th>Medial</th>
<th>Initial</th>
<th>Isolated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ب</td>
<td>بت</td>
<td>بیب</td>
<td>بی</td>
</tr>
<tr>
<td></td>
<td>ت</td>
<td>تیب</td>
<td>تا</td>
<td>تی</td>
</tr>
<tr>
<td></td>
<td>ث</td>
<td>ثب</td>
<td>ثب</td>
<td>ثب</td>
</tr>
<tr>
<td></td>
<td>س</td>
<td>صب</td>
<td>صب</td>
<td>صب</td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th>Urdu Writing Style</th>
<th>Forms of Non-Joiners</th>
<th>Final</th>
<th>Isolated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ان</td>
<td>یہ</td>
<td>د</td>
</tr>
</tbody>
</table>
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 abreide 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 \( Y = (Y_v)_{v \in V} \) 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: \( p (Y_v | X, Y_w, w \sim v) \) 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 models that are trained discriminatively. According to LDCRF let a given sequence of observations say, \( X = x_1, x_2, x_3, \ldots, x_n \) 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.

\[
P(Y/X) = \sum_h p(h/X)P(Y/h) \tag{1}
\]

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

\[
P(x_{1:n} | x_{1:n}) = \frac{1}{Z} \exp \left( \sum_{n=1}^{N} \sum_{i=1}^{F} \lambda_i f_i(z_{n-1}, z_n, z_{1:n}) \right) \tag{2}
\]

Where the normalization factor \( Z \) is calculated as under

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

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.

![NER model for segmentation.](image-url)
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.

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 (Shahi Qilla), if space is omitted then it will look like 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: (will do), (for), (narrate) (at that time), after omitting the space in between these words they make the forms: (أپ، گوگا، نیبیپ، گوگا) 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 (jail khana, Prison), مہنات و عظومت (mehnat o azmat, hardworking and greatness) and ذائرہ جواب (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 بند (din ba din, day by day), کبہ کبہ (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 پہلی ب (pahli b) word the morpheme ب (b) is added to the repeated word پہلی (pahli). The reduplicated words will treat by the system as separate orthographic words (Durani & Hussain, 2010).

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

References

For detailed information and references, please consult the original document.
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 ہد اخلاق (bad akhlaq, bad character), ہد وقاز (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), M.Phl (این فل) 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>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>U01:%x[-1,0]</td>
<td>N-1 token</td>
</tr>
<tr>
<td>U02:%x[0,0]</td>
<td>Current token</td>
</tr>
<tr>
<td>U03:%x[1,0]</td>
<td>N+1 token</td>
</tr>
<tr>
<td>U04:%x[-1,0]/%x[0,0]</td>
<td>N-1 word and N+1 token</td>
</tr>
<tr>
<td>U05:%x[0,0]/%x[1,0]</td>
<td>Current token and N+1 token</td>
</tr>
<tr>
<td>U06:%x[-1,0]/%x[1,0]</td>
<td>N-1 token and N+1 token</td>
</tr>
<tr>
<td>U07:%x[-1,1]</td>
<td>POS tag of N1 token</td>
</tr>
<tr>
<td>U08:%x[0,1]</td>
<td>POS tag of the current token</td>
</tr>
<tr>
<td>U09:%x[1,1]</td>
<td>POS tag of N+1 token</td>
</tr>
</tbody>
</table>

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
VIII. EXPERIMENTS AND EVALUATIONS

To evaluate the performance of our proposed system we used WordSeg 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 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>
<thead>
<tr>
<th>Tested Text</th>
<th>Sentence</th>
<th>Words</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>23</td>
<td>100%</td>
<td>50%</td>
<td>67.5%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>50</td>
<td>94%</td>
<td>51%</td>
<td>66%</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>99</td>
<td>94%</td>
<td>51%</td>
<td>66%</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>497</td>
<td>96%</td>
<td>50%</td>
<td>65.7%</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>Results Comparison of Proposed CRF Model with Baseline Approach</th>
<th>Tested Prov</th>
<th>Correctly Segme</th>
<th>Uncorrecte</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach</td>
<td>Problem Addressed</td>
<td>Text</td>
<td>Segmented Words</td>
<td>Segmented Words</td>
</tr>
<tr>
<td>Baseline</td>
<td>Space Omission</td>
<td>11.99</td>
<td>11.723</td>
<td>272</td>
</tr>
<tr>
<td>Proposed CRF Approach</td>
<td>Space Omission, Deletion, compound</td>
<td>3,161</td>
<td>3,118</td>
<td>43</td>
</tr>
</tbody>
</table>

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

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


