Development and Enhancement of Tools and Resources for Urdu Text Processing

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science in **Computational Linguistics** by Research*

by

Aamir Farhan
20161078
aamir.farhan@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA
April 2023

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **"Development and Enhancement of Tools and Resources for Urdu Text Processing"** by **Aamir Farhan**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. Dipti Misra Sharma

To My Beloved Family

Acknowledgments

First and foremost, I would like to thank the Almighty, who is the origin of all knowledge in this world and who bestows as much as He chooses onto everyone He deems acceptable. My deepest gratitude goes out to my advisor Dr. Dipti Misra Sharma for her constant and unparalleled guidance, help and support throughout my time at IIIT Hyderabad. I'll forever be grateful to her for instilling the linguistic curiosity and passion for research in me. I couldn't have hoped for a finer mentor to help me with my research, academics and life in general.

Besides my advisor, I would also like to express my gratitude towards my professors who taught and guided me over the years, especially Dr. Rajeev Sangal, Dr. Radhika Mamidi, Dr. Manish Shrivastava, Dr. Harjinder Singh and Dr. Aruna Chaluvadi. The courses and projects I studied with them have worked immensely towards my holistic growth and development. It has been a privilege for me to have gotten a chance to study under them. I would also like to thank Arafat Ahsan, Riyaz Ahmad Bhat, Irshad Ahmad Bhat, Pruthwik Mishra, Vandan Mujadia and Rashid Sir for providing enormous help and guidance throughout my research journey. Their expertise in the subject has played a crucial role for me in tackling the hurdles and challenges I faced during the research work. The journey seemed smooth despite of all the ups and downs because of them.

I express my sincere gratitude towards my batch UG2k16 for constantly showing faith in me and also electing me as their Felicity coordinator. All those fun times played an important role for me to take time off from the hectic schedule and grow holistically. All of my batchmates, especially Ali, Arpita, Mannan, Ganesh, Keshav, Yaseen, Pravin, Sourav and Abheet, thanks to all of you!

Finally, I would like to thank my beloved family for their endless love and support. It is because of their trust in me, well wishes and everyday prayers that kept me going and stay motivated in this journey.

Thank you, everyone!

Abstract

Urdu writing system is derived from the Persio-Arabic writing systems and thus it has adopted similar orthographical and morphological characteristics as that of Persio-Arabic languages. The first and foremost task for most of the NLP applications is Word Segmentation which involves identifying the bounding boundaries of words in written text. It is quite crucial to accurately identify the boundaries of each word in written text because all the downstream tasks in NLP are dependent on it, thus making Word Segmentation fundamentally important. Urdu adopts a continuous writing style which does not have an explicit and clear marker for word boundary. Furthermore, the inherent non-joining attributes of certain characters in Urdu create spaces within a word while writing in digital format. Thus, Urdu not only has space omission but also space insertion issues which make the word segmentation task challenging. We have studied and categorized the various issues that are observed with respect to the inconsistent usage of space character in Urdu script along with the orthographic and morphological reason behind it. Another challenge in computational processing of Urdu is the lack of benchmark resources and corpora for Word boundary identification.

Leveraging the learning from the orthographic study of Urdu writing system, we have built a benchmark corpus for Urdu Word Segmentation, with an exercise of manual annotation, using white space as word boundary and Zero-Width-Non-Joiner (ZWNJ) character as sub-word boundary. A Conditional Random Field based sequence modeler was then used to train a character-level label prediction of a sequence of Urdu characters. Our model achieved state-of-the-art results with an F_1 score of 0.98 for word boundary identification. Furthermore, we have applied our word segmentation model on studying the sociological phenomena of Diglossia in Urdu.

Contents

List of Figures

List of Tables

Table Page

Chapter 1

Introduction

Urdu is an Indo-Aryan language and one of the 22 official languages of India. It is widely spoken in South Asian countries and as a result of significant South Asian diaspora, it is also spoken all over the world with an estimate of 160 million speakers worldwide[1], either as their mother tongue or as their second language. Urdu is a morphologically rich language which uses Arabic based orthography. Morphology of Urdu language is influenced by many languages including Arabic, Persian, Turkish, Hindi, Sanskrit and English. (Waqas et al., 2006) (Riaz, 2007).

Computational processing of text can be challenging for languages which do not have a clear delimiter to mark word boundary in their writing system. Urdu is one such language which faces complex issues in word segmentation because of its rich morphology and conventional way of writing it in a fashion which gives a lot of importance to write correct shape of letters based on their relative position in a word. So, due to non-availability of an explicit delimiter and its orthography and rich morphology, Urdu faces complex and interesting challenges in word segmentation.

1.1 Motivation

The first and foremost task for all natural language processing systems such as part-of-speech tagging, machine translation, information retrieval and extraction, question-answering systems, grammar checker etc. is word segmentation. Word segmentation is a process of finding the word boundaries in written or typed text. Word boundaries or word segments must be clearly identified in the input text for each of the above mentioned language processing systems so that all the downstream tasks can be performed. Urdu Word Segmentation is a challenging task because of numerous issues including space insertion and space omission, which are discussed in detail in Chapter 2 and Chapter 3. Since, identifying word boundaries is the first and crucial step for all Urdu text

processing applications, it becomes quite important to build an accurate, robust and efficient system which performs the word segmentation. Moreover, as is the case with all machine learning and deep learning architectures, their models require large amounts of training data which is a barrier for low-resource languages like Urdu. Therefore, development of standard and good quality resources and datasets is also essential so that they can be leveraged for various NLP applications.

1.2 Background and Related Work

For various languages around the world, word segmentation problems have been solved using a variety of strategies. For Chinese word segmentation, many rule-based techniques were adopted which creates all potential segmentations of letter sequences before choosing the most effective one using heuristics like long word lengths, minimum number of words etc. (Sornlertlamvanich, 1993; Ping and Yu-Hang, 1994; Nie et al., 1994). Machine Learning methods are also quite popular in identifying the word boundaries in a sentence. These methods make use of learning algorithms that can define a function that takes input samples and outputs a variety of values. For these methods, a corpus is created where word boundaries are well established. There are statistical models created that include characteristics of the boundaries-enclosed words. In NLP, supervised statistical learning is currently one of the most used techniques. (Wu et al., 1994) (Kaplan, 2005) (Xing et al., 2008). For effective word segmentation, Cai and Zhao (2016) and Cai et al. (2017) studied the application of neural language models with word and character based embedding and outperformed conventional segmentation techniques.

Durrani and Hussain (2010) have conducted significant research on the segmentation of words in the Urdu language, utilizing hybrid techniques including rule-based methods that rely on maximum match, statistical techniques like n-grams, and word POS tags. For space insertion and space omission issues, their best techniques had an error detection rate of 85.8%. More recently, Zia et al. (2018) proposed a CRF based model for Urdu Word Segmentation along with a publicly available corpus which showed better results than Durrani and Hussain's model. We have outperformed the results of Zia et al. by using a manually annotated, much larger corpus of 19,651 Urdu sentences along with a more optimal set of contextual features.

1.3 Problem Statement and Contributions

As part of the research on the topic of this thesis, we have done an intensive and thorough study and analysis of the

Urdu's writing system, orthography, morphology and sociology. We have identified and classified the numerous writing rules and measures which give rise to various text processing and segmentation challenges. The major problem that we have solved as part of this research is the Word Segmentation problem in Urdu which involves identifying word peripheries in written Urdu by handling the unique and fundamental issues which make this task quite challenging and critical, as discussed in Chapter 3.

The key contributions of this research work are:

- Manually annotated benchmark corpus for Urdu Word Segmentation task, which is the biggest in terms of number of sentences so far.
- State-of-the-art Urdu word tokenization model with optimally crafted feature set.
- First of its kind handling and annotation of special grammatical constructions in Urdu such as *Izafa* constructions.
- A social study and analysis of Urdu language which validates the existence of Diglossic situation of Urdu in South-Asian countries.
- A rich and extensive lexicon of Urdu tokens with approximately 200,000 accurately segmented words.

1.4 Publication

The research work majorly described in Chapter 2, Chapter 3 and Chapter 5 of this thesis discussing the corpus development and model training for Urdu Word Segmentation has been presented as a publication, being the first author as:

- **Aamir Farhan,** Mashrukh Islam, and Dipti Misra Sharma. "Enhanced Urdu Word Segmentation using Conditional Random Fields and Morphological Context Features" In Proceedings of the ACL (2020) workshop in collaboration with The Fourth Widening Natural Language Processing Workshop.

1.5 Thesis Overview

- **Chapter 2**. In this chapter, we discuss the orthography of Urdu and its writing system. We describe various types of joining forms and each Urdu character attains based on its relative position in the word. This study and analysis is necessary because the orthography forms the foundation of the challenges that we encounter in Urdu text processing.

- **Chapter 3.** This chapter is dedicated in identifying and classifying the various space related issues that are observed in the script which make the word segmentation in Urdu a unique case. We have discussed all such space insertion and space omission issues in detail in this Chapter.

- **Chapter 4.** In this chapter we have shown our preliminary work in solving the problem at hand using a rule based system. The system presented in this chapter works on the algorithm of lexicon based maximum matching heuristic.

- **Chapter 5.** This chapter focuses on our approach of solving the word segmentation problem using machine learning concepts and our journey of developing the training corpus for this task. We have discussed in detail, the numerous scenarios and constructions we have handled while corpus development. We have also discussed in depth the feature engineering done with different combination of features in order to get the optimal precision, recall and f1 score.

- **Chapter 6.** In this Chapter we have discussed the phenomena of Diglossia in context of Urdu speaking spaces. Our analysis, using our in-house word segmentation tool, validates the existence of Diglossic situation for the language Urdu.

- **Chapter 7.** We offer our final conclusions and lay out potential avenues for future research on Urdu text processing in this chapter.

Chapter 2

Urdu Orthography

Urdu is an Indo-Aryan language widely spoken in India, Pakistan and Bangladesh. Urdu alphabet is a variant of the Persian alphabet, which itself is a descendant of the Arabic alphabet. The Urdu script consists of 37 basic and four secondary letters, seven diacritics, punctuation marks, and special symbols in the character set. Urdu shares a common script and many characteristics of Arabic script with additional set of alphabets to cover its much wider repertoire of sounds as shown in the appendix. Urdu orthography or the Urdu writing system shares several traits with Arabic orthography, like the optional usage of diacritic markings.

2.1 Bidirectional Urdu Script

Urdu text is written and read in horizontal lines from right to left direction, while numbers and ingrained Latin text are written and read from left to right. Thus, Urdu script is said to be bidirectional in nature in the setting of reading and writing as shown in *Figure 2.1* . Urdu is written in an Arabic cursive writing system. Urdu characters, in general, unite with their neighbors within a word and take on distinct shapes as a result.

ہندوستان ۱۵ اگست ۱۹٤۷ کو آزاد ہوا تھا

English Gloss : *hindustan 15 august 1947 ko azaad hua tha*

English translation : India got independence on 15 August 1947

Figure 2.1 Bidirectionality in Urdu Script

2.2 Diacritics and Omission of Short Vowels

Urdu writing style is context sensitive by nature, which means that characters change shape depending on the characters that come before and after them. Diacritics are employed to ensure appropriate pronunciation of constituent words. Diacritics can be found either above or below a character as shown in *Figure 2.2*
The Urdu script only depicts consonants and long vowels while short vowels are often omitted while writing except when needed to distinguish between two words as shown in *Figure 2.3,* the word تیر , which is written without diacritics is ambiguous without context which can be interpreted as *taer* (swim) or *teer* (arrow).

The absence of short-vowel symbols, on the other hand, does not pose a problem for readers because words exist in a linguistic context, and the word immediately before or following one may be all the reader needs to comprehend the correct interpretation. Any machine transliteration or text-to-speech synthesis system must guess and insert the missing symbols automatically. This is a difficult subject that necessitates a thorough statistical examination.

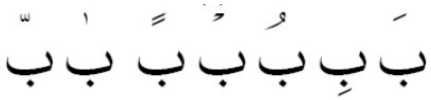

Figure 2.2 Location of diacritics

Without Diacritics	**Multiple Ambiguous Meanings**	**With Diacritics**	**Unambiguous Meaning**
تیر	Swim (taer) Arrow (teer)	**تِیر**	Arrow (teer)
بل	Hair (bal) Invoice (Bill)	**بِل**	Invoice (bill)
جھیلو	Undergo (jhelo) Lakes (jheelo)	**جھِیلو**	Lakes (jheelo)

Figure 2.3 Urdu words with and without diacritics

2.3 Joiners and Non-Joiners in Urdu

Each letter in the Urdu script, depending on its position in the word, can attain one of its numerous forms. When a letter is written alone, it is drawn in one form, and when it is joined to other letters in a word, it is written in up to three other forms. Based on their position relative to a word, characters can attain one of the following forms : initial, medial, final and isolated. Urdu characters can broadly be divided into joiners and non-joiners depending upon the number of positional forms they can attain.

Joiners are characters that can take all four shapes, whereas non-joiners are characters that only have two forms: final and isolated. So, Non-joiners are characters who only join with characters which come before them but not with characters which come after them. Table 2.1 shows Urdu joiners and non-joiners in their isolated form

Non-joiners	د ا ڈ ذ ر ڑ ز ژ و ء ے
Joiners	ب پ ت ٹ ث ج چ ح خ س ش ص ض ط ظ ع غ ف ق ک گ ل م ن ں ی

Table 2.1 Joiners and Non-joiners in their isolated form

2.4 Orthographic Rules of Joining

We distinguished and identified the following shape changes in the characters when a joiner or non-joiner is followed by joiner or non-joiner.

Case 1 : Joiner in the initial position

The character م takes its initial form when it is the starting character of a word.

For example, مومکین , *mumkin* (possible)

Case 2 : Joiner after joiner in the middle position

The character م takes its medial form when it is written after another joiner in a word.

For example, نمکین , *namkeen* (salty)

Case 3 : Joiner after non-joiner in the middle position

The character م takes its initial form when it is written after a non-joiner in a word.

For example, دماگ , *dimaag* (brain)

Case 4 : Joiner after joiner in the final position

The character م takes its final form when it is written after another joiner in the final position in a word.

For example, رسم , *rasm* (ritual)

Case 5 : Joiner after non-joiner in the final position

The character م takes its isolated form when it is written after a non-joiner in the final position in a word.

For example, شام , *shaam* (evening)

Case 6 : Non-joiner in the initial position

The character د takes its isolated form when it is the starting character of a word.

For example, دنیا , *duniya* (world)

Case 7 : Non-joiner after joiner in the middle position

The character د takes its final form when it is the middle character written after a joiner in a word. For example, بندک , *bandook* (gun)

Case 8 : Non-joiner after non-joiner

The character د takes its isolated form when it is written after a non-joiner in a word.

For example, فریاد , *fari*yad (complaint)

As observed in the above scenarios, we come to the conclusion that the following three factors deterministically decide the shape of a character in a word:

1. The inherent property of being a joiner or non-joiners.
2. The relative position of the character in a word (initial, medial, final or isolated)
3. Whether the character written immediately before the character of interest is a joiner or a non-joiner.

The above observations and patterns were studied over hundreds of sentences written in Urdu derived primarily from Urdu news websites like BBC Urdu. We have utilized these rules and patters in modeling the features required for training the word segmentation model described in Chapter 5.

Chapter 3

Word Segmentation in Urdu : A Unique Case

The process of word segmentation involves locating word boundaries or word bounding boxes in written text. In this process, the text is split into its individual and discrete words (Kaplan, 2005). (Manning et al., 1999). All language processing systems, including machine translation, part-of-speech tagging, information retrieval, information extraction, grammar and spell checkers, do this preliminary work of identifying word peripheries. Word boundaries must be clearly defined in the input text for each of these language processing systems. In this chapter, we will describe the challenges faced in word boundary identification in Urdu and why this task makes a unique case for the language Urdu.

When it regards to the orthography, morphology and character types, Urdu is a complex language, as shown in Chapter 2. There is no traditional usage of white space to separate words in handwritten Urdu language. The word boundary is determined by the native speaker of the language simply by glancing at the character shapes. When there is a deterministic space between words, word segmentation is made easier, but in computer-typed Urdu text, the usage of space is incredibly uneven and is only used in certain circumstances, which makes the boundary identification much more difficult (Lehal, 2010). We have categorized all the challenges and inconsistencies with regards to irregular usage of space in the following sections.

3.1 Space Omission Problem

We described the orthographic rules of joiner and non-joiner characters of Urdu script in Chapter 2. In case of words that end with a non-joiner character, the right shape of the word is produced regardless of space typed after it. This is because the non-joiner character retains the same shape

with or without space. As a result, frequently, users don't add the space and write in a continuous fashion. Consequently, by not typing a space, the current word agglutinates with the next word and thus becomes a challenge for computational processing. Figure 3.1 shows a phrase in which the ending character of each word is a non-joiner and thus the entire phrase can be represented without space and is still visibly readable.

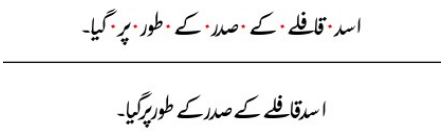

Figure 3.1 Urdu phrase written with and without spaces

English Gloss : *asad kafile ke sadar ke taur par gaya*

3.2 Space Insertion Problem

Space insertion problem is faced when ending characters of Urdu morphemes are joiners. Many of the morphemes that are written together within a word tend to keep their own ligature shapes. Thus, writers typically include space to stop them from joining and preserve the distinct orthographic identity. For example, Urdu typists learn to insert a white space within the word خوش قسمت (Fortunate) to get the correct shape of ش. Without space, it appears like خوشکسمت which is visually incorrect.

As a side-effect of producing the visually correct form of the character, this actually splits a single word into multiple tokens during processing. These multiple morphemes within a word would merge if the writers did not type a space between them, creating an incorrect shape and ligature. The problem of space insertion is found to be prominent in scenarios of Affixation, Compounding, Reduplication, Foreign word and Abbreviation. These issues have been individually discussed across various scattered sources [2], [4], [8]. We have thoroughly explored and studied the fragmented literature for the issues and have categorized them as described below

3.2.1 Affixation

In Urdu, affixes exist as both suffixes and prefixes. A space character is put between the prefix and the stem whenever a prefix or stem is a separate morpheme and the prefix ends in a joiner. Similar to this, space is added between the stem and suffix if the stem terminates in a joiner. They must, however, be ideally contained within a single periphery because they are single semantic units.

Figure 3.2 With and Without Space for the word Khush-Naseeb (Fortunate)

3.2.2 Foreign Words

In Urdu language, a few English words are borrowed quite often. These words frequently consist of multiple morphemes. Space is inserted between these morphemes when the first one, when written in Urdu, ends with a joiner character. The tokenizer should disregard this space and give each of these words should be treated as a single semantic unit.

Figure 3.3 With and Without Space for the word Football

3.2.3 Acronyms and Abbreviations

In Urdu, English acronyms and abbreviations are written with spaces between the pronunciation of each character and are spoken like English characters written in Urdu. These acronyms and abbreviations function as a single token or segment. If any name is placed after these, they come together to form a single entity (Sproat, 1992).

Figure 3.4 With and Without Space for the word PhD

3.2.4 Reduplication

Words are often reduplicated in Urdu for semantic emphasis of an entity or event. These reduplicated words must be treated as a single token or segment. When one morpheme of these reduplicated words ends with a joiner, a space character is inserted to prevent it from joining and thus keeping its correct visual form.

Figure 3.5 With and Without Space for the *Dhoom-Dhaam* (Pageantry)

3.2.5 Compound Words

Compound words in Urdu are formed when two roots morphemes combine to create a single word with a meaningful unit. When the first morpheme in a compound unit ends in a non-joiner, a space is often neglected between them, as in محنت‌مشقت (hard work) . However, if the first

morpheme ends in a joiner, a space is added after it to get the correct visual representation. These compound words must be treated as a single segment and should be given a single word boundary (Sproat, 1992).

3.2.6 Izafa Constructions

Izafa constructions are special grammatical constructions found in Urdu and Persio-Arabic languages which are of form A-e-B. These constructions signify the possessive character and the -e- can be treated semantically similar to ‘of’ in English. In the A-e-B formation, two roots or stems are connected to one another by a linking morpheme which is denoted by the diacritic “ِ”. These morphemes combine to produce a single semantic unit. As discussed in Chapter 3, these diacritics are often removed while writing.

For example,

وزیر اعظم (Wazir-e-Azam)

There is no need to place a space between the first morpheme and the linking morpheme if the first morpheme ends at a non joiner. However, space is inserted between the first morpheme and the linking morpheme if the first morpheme ends in a joiner. From computational point of view, this sort of special constructions must be given a single word boundary, with the inter word gap removed. We will discuss more about Izafa constructions in Chapter 5.

3.3 Impact of Space related issues on Word Segmentation

According to an early study conducted on the subject by Naseem and Hussain (2007), a considerable amount of improper usage of space was observed , as shown in the above sections.

In this study, an Urdu corpus was checked for its correctness with respect to the spelling of words. The checker's reported errors were manually examined. A dataset of approximately two thousand Urdu sentences was analzed. A total of 975 faults were discovered, of which 736 (75.5%) were attributable to the erroneous use of space, and 239 (24.5%) were attributed to human errors. The majority of space-related errors—672 or 70% of all errors—are caused by space omission, whereas 53 errors (or 5% of all errors) were caused by space insertion.

Because irregular usage of space results in a very high percentage of errors, it makes the regularization and proper word segmentation of Urdu an extremely important task.

Another study was conducted by Durrani and Hussain (2010) in which popular Urdu online news sources (BBC Urdu and Jang.com) were used for the Urdu text. A dataset of five thousand Urdu words from each corpus were manually analyzed, and instances of space insertion and omission were counted as shown in Table below.

Problem	BBC	Jang	Total
Space Omission	373	563	936
Space Insertion			
Affixation	298	467	765
Reduplication	52	76	128
Compounding	133	218	351
Abbreviation	263	199	462
Total	1119	1523	2642

Table 3.1 Space Omission and Insertion Counts

Chapter 4

Rule Based System for Urdu Word Segmentation

In the previous chapters we showed that space omission and space insertion are major and frequent issues in Urdu text because of the incapability of the modern Urdu typing technologies in adapting the orthographic nature of the script and the connecting nature of Urdu letters. Thus, to correctly identify the word boundaries in Urdu text and address space-related challenges, an effective Urdu word segmentation system is required. To tackle this challenge, we first adopted an unsupervised rule-based approach using dictionary lookup and maximum matching algorithm. Later on, we moved on to a deep learning approach using Conditional Random Field to train a model on an in-house annotated corpus for the word boundary identification, as discussed in Chapter 5. In the current Chapter, we will describe our preliminary approach to solve the problem in hand and its drawbacks.

4.1 Maximum Matching Dictionary Look-up

In general, there are three kinds of dictionary look-up based maximum matching algorithms viz. forward maximum matching, backward maximum matching, and bidirectional matching. We adopted the forward maximum matching algorithm with a dictionary of approximately 200,000 Urdu words.

4.1.1 Max-match Algorithm

Lets say that D is the name of the global dictionary that contains all the entries of Urdu words. The name of

the longest entry is MaxLen. Consider the Urdu character string S = (s1, s2, , sM) to be of length M. The forward maximum matching technique takes the candidate string of the phrase from right to left in order to find the boundaries of the words in the statement.

$$k = max_h hI((s_1, s_2, \cdots, s_h) \in D.$$

Where

$$I((s_1, s_2, \cdots, s_h) \in D) \quad = \quad \left\{ \begin{array}{l} 1, (s_1, s_2, \cdots, s_h) \in D \\ 0, (s_1, s_2, \cdots, s_h) \notin D \end{array} \right\}$$

Here, (s1, s2, · · · , s_h) denotes a matching hit with one of the entries in the global dictionary. Therefore, (s1, s2, · · · , s_h) will be segmented out as a word. The algorithm then proceeds to find the next maximum. The segmented word is referred to as a correct entry when k = 1. When k = 0, the segmented word is referred to as an unregistered word.

The backward maximum matching algorithm and the forward maximum matching algorithm have the same fundamental idea. It is solely used for word segmentation of sentence in reverse. Figure 4.1 shows the flow diagram of the steps involved in the dictionary based maximum matching algorithm.

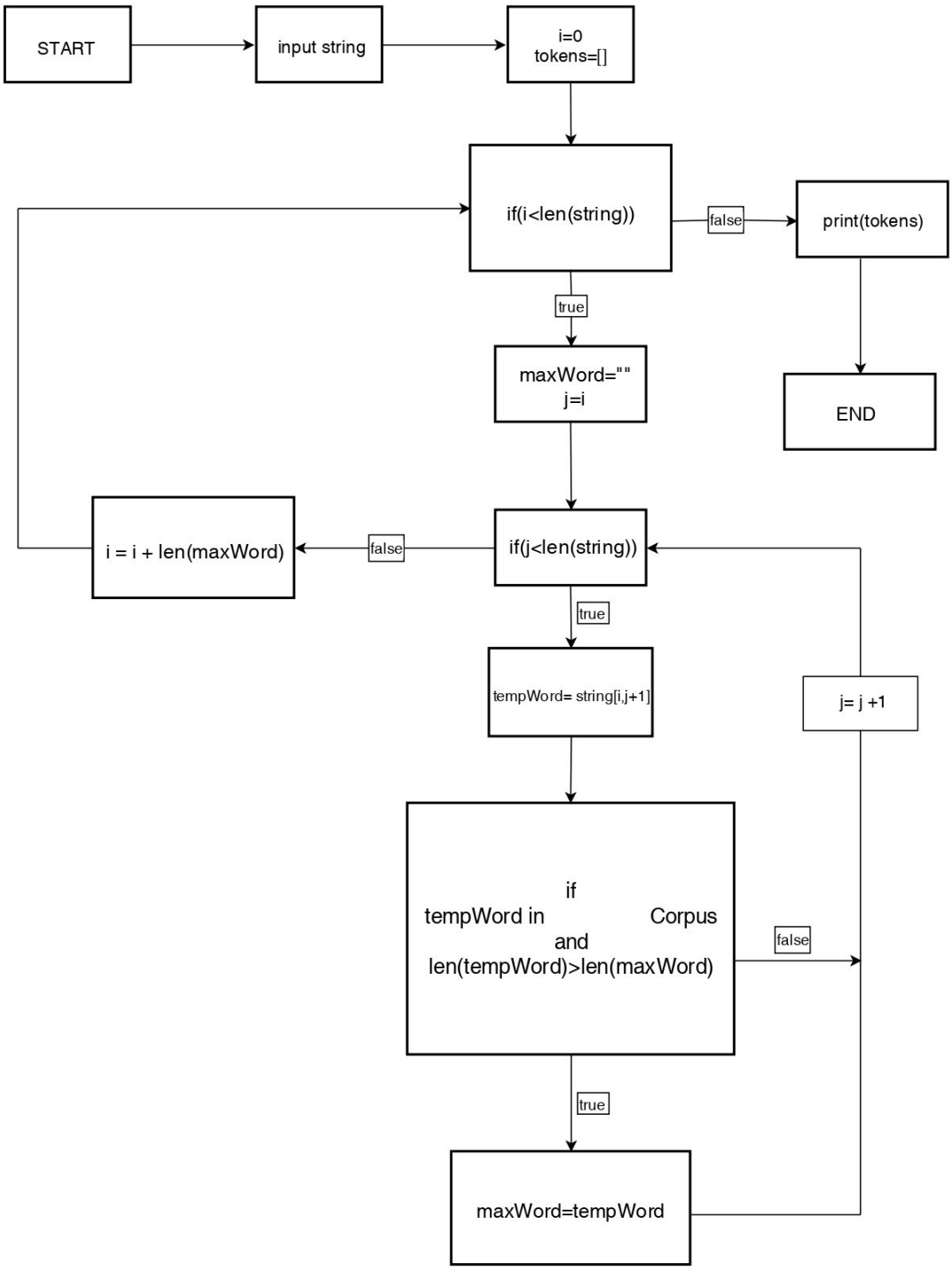

Figure 4.1 Flow diagram depicting dictionary based maximum-matching algorithm

Lets take, for example, the following input string in Urdu

مشینونکےرکھرکھاؤصحیحنہینہے

English Transliteration : *masheenon ke rakh rakhaav sahee naheen he*

English Translation : maintenance of machines is not right

After applying the dictionary based max-match algorithm we get the following segmentation of words separated by a space

مشینون کے رکھ رکھاؤ صحیح نہین ہے

4.2 Limitations and Drawbacks

The dictionary based max-match algorithm described in this chapter comes with various drawbacks and challenges. Firstly, the algorithm does not take context of the sentence into account. This implies that the capability to detect ambiguity and selecting contextually best segmentation is lacking in the algorithm. As a consequence,

1) Shorter word sequences are never generated even when they are intended.
2) The max-match heuristic fails when alternatives have same number of words.

For example, considering the input string

کھاناکھالیا

English Transliteration : *khaanaakhaaliyaa*

English Translation : Had meal

This string will get segmented into ***khaanaa khaaliy aa*** (incorrect segmentation) instead of ***khaanaa khaa liya*** *(correct segmentation).*

Moreover, the conventional max-match word segmentation method uses the global dictionary to divide the string into segments in accordance with a certain pattern. This makes the word segmentation outcome more influenced by the dictionary's level of granularity. In Chapter 5, we present a more robust and highly accurate machine learning approach to identify the correct word segmentation of an Urdu string.

Chapter 5

Word Segmentation as a Sequence Labeling Task

Sequence labeling is a fundamental pattern recognition task in NLP that surrounds a wide range of tasks, such as named entity recognition (NER), POS tagging, text chunking, and so on. The sequence labeling task in the NLP applications can be defined as a task that seeks to assign labels to a group of meaningful units in a sentence that play similar roles in sentences' grammatical structure and have similar grammatical features. The type of labels and their meanings are usually determined by the type of task involved.

In general, traditional sequence labeling algorithms rely significantly on hand-crafted features or language-specific resources, and are based on classical machine learning technologies like Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs). To achieve superior performance using these techniques, it is required to have considerable amount of domain knowledge and efforts on feature engineering.

5.1 Markov Assumption

The majority of sequence labeling algorithms are probabilistic, depending on statistical inference to determine the optimum sequence. A Markov assumption is used in the most frequent statistical models for sequence labeling. In Bayesian probability theory, the Markov assumption states that every node in a Bayesian network is conditionally independent of its non-descendants, given its parents. In other words, the label for a given word is solely determined by the labels immediately

adjacent to it. Thus, all the previous labels can be ignored and the probability of the next label prediction can be approximated to depend only on the probability of the few adjacent labels as shown below:

$$P(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_n = x_n \mid X_{n-1} = x_{n-1}).$$

The collection of labels are said to form a Markov chain when the label for a given sequence is solely determined by the labels immediately adjacent to it. The most frequently used statistical models for sequence labeling that make a Markov assumption are Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs). We have utilized a CRF based architecture for training the model for word segmentation in Urdu which is described in detail in the upcoming sections.

5.2 Character-wise Sequence Labeling

To model the given problem of Urdu word segmentation as a sequence labeling task, we adopted a character level labeling scheme. In this labeling scheme, each character of a sequence of Urdu text is assigned a label as shown in Table 5.1

Label	Annotation
B	Beginning character of a word
S	Beginning character of a sub-word
O	Other characters

Table 5.1 Label set for character-wise annotation

For example, in the sequence خوش‌قسمت (*khush-kismat)* the character-wise labeling would result in the following label set [B,O,O,S,O,O,O].

5.3 Corpus Development

When we look into the benchmark corpora of various languages to perform fundamental text processing tasks, we find that Urdu is highly deficient in resources. Moreover, the existing resources and corpora for Urdu Word Segmentation lack the dimension of quantity and hence, are too small to encompass several morphological construction that exist in the Urdu language. To overcome this shortcoming, we manually annotated and developed a benchmark corpus for Urdu Word Segmentation.

5.3.1 Data Collection

As a starting step, raw data of Urdu text was collected from several online sources such as BBC Urdu, Siasat Daily, Urdu Point. The gathered raw data is free and openly available to public for research purposes. To make sure that the collected data encompasses various morphological constructions such as affixation, reduplication, compounding, abbreviation and foreign word, the data was gathered from a variety of genres and domains such as Entertainment, Commerce, Weather, Sports, Health, Politics and Technology. After considering each of these domains uniformly, a total number of 19,651 sentences were finally selected to start the annotation process.

5.3.2 Data Annotation

The major markers that we incorporated in the annotation process are as follows:

We used a white space character to mark the word boundaries. This means that a white space decides the strict breaking point for a word or a token. We used a **Zero-Width-Non-Joiner or ZWNJ** character to mark the sub-word boundary. The advantage of using a ZWNJ character is that it prevents a joiner character in Urdu from taking its “joiner-form” and thus not letting it connect with the subsequent

character and at the same time retaining its visually correct form. For example, the word *khush-naseeb* (Fortunate), with and without a ZWNJ, would look like shown in Table 5.2:

خوش‌نصیب	خوش نصیب
With ZWNJ (Correct Form)	Without ZWNJ (Incorrect Form)

Table 5.2 With and without ZWNJ

While annotating the collected data, we followed the rules proposed by Rehman et al. (2011) as shown below

- For marking the precise boundary of a word as a token, white space was used to separate out words. The incorrect usage of white space compared to correct usage is shown in Table 5.3

اسد شہر سے باہر جا پہنچا (II)	اسدشہرسےباہرجاپہنچا (I)
Asad reached out of the city.	

Table 5.3

- Compound words of form **AB** in Urdu are often written with a white space between them as shown in Table 5.4 below. To get the correct shape of compound words, a Zero-Width-Non-Joiner is inserted between two roots of a compound word.

محنت مشقت (hard work)
روٹی کپڑا(basic needs of life)
ماں باپ (parents)

Table 5.4 White space usage in compound words

- For getting the correct formation of special constructions like prefixes, suffixes and reduplication, a Zero-Width-Non-Joiner is inserted as shown in Table 5.5

Construction	Usage of ZWNJ
Prefixation	خوش‌نصیب (Fortunate)
Suffixation	غلط‌فہمی (Misunderstanding)
Reduplication	صبح‌صبح (Early Morning)

Table 5.5 Prefixation, Suffixation and Reduplication with ZWNJ character

- In scenarios when English words are transliterated and used in the Urdu script, these words consist of multiple morphemes. When the ending character of these morphemes is a joiner, a ZWNJ character is inserted to prevent it from joining. For example, نیٹ‌ورک (network) .

- When English acronyms and abbreviations are written in Urdu script, they behave as a single word. Thus, a ZWNJ is inserted in between the acronym and abbreviation units to prevent them from joining. For example, این ایل پی (NLP).

5.3.3 Izafa Constructions

In Urdu, *Izafa* is a special syntactical structure consisting of two nouns, the first of which is a determined noun and the second of which is a determiner. Inspired from Persio-Arabic scripts, Izafa constructions in Urdu are usually used to denote possession. In Urdu, an unstressed short vowel '-e-' is used to connect the two syllables, and the short vowel is attached to the first word when pronouncing the newly created word, but it is rarely written in the script as shown in Figure 5.1.

وزیر اعظم (prime minister)
طالب علم (student)
حد نظر (scene limit)

Figure 5.1 Izafa Constructions in Urdu

This short vowel is frequently used in a way that is similar to the English preposition 'of.' As seen in previous sections, Urdu script is normally written without the short vowels, thus the Izafa vowel ‘-e-’ is not indicated in written Urdu script. Hence, identifying these special constructions and annotating them as a single unit becomes an important exercise in the word segmentation model.

In our process of annotation, we assign a single word boundary to Izafa constructions by replacing the space between two words with a Zero-Width-Non-Joiner character. For example, وزیراعظم
To do further morphological analysis of Izafas, we propose that these constructions should also be assigned a label indicating their behavior and thus can be broken down into its morphemes as required.

5.3.4 Corpus Analysis

Following the rules and observations as discussed in section 2.3.2 and section 2.3.3 of Chapter 2, a total number of 19,651 sentences were annotated by two annotators who are native Hindustani speakers with a background in Computational Linguistics and knowledge of Urdu script. This corpus is significantly larger when compared to previous work done on Urdu Word Segmentation, particularly by Zia et al. (2018), which presented a corpus of 4,325 sentences. We divided the data into the following sets of sentences for evaluation purposes: 17,401 sentences for training and 2250 sentences for testing.

To perform a qualitative analysis of our corpus, we conducted an inter-annotator agreement study in which a total of 100 sentences were exchanged and annotated by the two annotators to determine the inter-annotator agreement. A good inter-annotator agreement ensures that the annotations are reliable. Using the Cohen’s Kappa coefficient, the inter-annotator agreement was found to be 0.96. For categorical data, Cohen’s Kappa is the most used annotation agreement coefficient.

5.4 Feature Crafting and Experimentation

Traditional sequence labeling algorithms, as previously mentioned, rely heavily on hand-crafted features or language-specific resources to achieve optimal performance. To get good training outcomes, it is required to have considerable amount of domain knowledge and significant effort put into feature engineering. Zia et al. (2018) presented a set of different linguistic and orthographic features for training the model. We experimented with those features and crafted our own set of new features which gave us better performance in terms of f1 score, precision and recall. Table 5.8 summarises the impact of each feature on the model's performance. The final set of features which made it to the model training and gave the optimal results are described below

- **4-grams :** For a given character in Urdu script, an incremental window of up to 4 neighbouring characters of either sides are considered. For example, the character **م** in the phrase**لگاکر الزا م** will have the set up to 4-grams as part of its features as shown in Figure 5.2:

['c=م', 'c-ا', 'c-ام', 'c-زام', 'c-لزام', 'c-الزام', 'c+ل', 'c+مل', 'c+م' لگ, 'c+م' لگا, 'c+م' لگاک]

Figure 5.2 4-gram character window; c= refers to the current character; c- refers to previous characters; c+ refers to next characters

- **Window between tri-gram and 4-gram :** This feature is similar to the 4-gram feature described above. Here, we only consider the characters between the third neighbour and the fourth neighbour of the current character on either sides. For example, for n^{th} character x_n, the sequence window **[x_{n-4} , x_{n-3} , x_n , x_{n+3} , x_{n+4}]** is taken as a feature.

- **Type of Character :** A boolean feature marking whether the current character is a joiner or a non-joiner. Refer to section 2.3 for a comprehensive list of the joiners and non-joiners in the Urdu script.

- **Consonant or Vowel :** A boolean feature marking whether the current character is a consonant or a vowel of Urdu script.

- **Bidirectionality :** As discussed in Chapter 2, in Urdu script the digits follow left to right direction whereas the characters follow right to left direction. This is also a boolean feature which marks the directionality of the current character.

- **Character Frequency :** A normalized number denoting the frequency of the current character in the script denoting how frequently the character of interest can come at a position of word boundary or sub-word boundary.

We also experimented with increasing the n-gram window for the feature. However, after the window size of 4-gram, there was no significant improvement in the performance of the model and hence the n-gram features with n value greater than 4 did not make it to the final set of features used for training the model.

5.5 Conditional Random Field Architecture

In many NLP applications, the task of assigning label sequences to a group of data sequences arises. Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) are the most prominent approaches for performing such labeling and segmentation tasks. Conditional Random Field models are a probabilistic framework used to label a novel data sequence **X** by choosing the label sequence **Y** with the highest conditional probability **p(Y|X)**. For a given data sequence, a CRF is an undirected graphical model that specifies a single log-linear distribution over label sequences.

CRFs offer increased flexibility relative to HMMs for sequence labeling because rather than modeling joint probabilities over observations and corresponding labels, CRFs model conditional probabilities over label sequences given specific sequences of observations. This allows CRFs to more easily capture long range dependencies and helps them avoid making inaccurate independent assumptions. CRFs also overcome the label bias issue, which is a flaw in maximum entropy Markov models (MEMMs) and other conditional Markov models based on directed graphical models. On a variety of real-world sequence labelling tasks, CRFs beat both MEMMs and HMMs.

5.5.1 Undirected Graphical Model

Conventionally, we define G = (V, E) as an undirected graph in which each of the random variables representing an element Y_v of Y has a node $v \in V$ corresponding to it.
(Y, X) is a conditional random field if each random variable Y_v obeys the Markov property with regard to G. We have described Markov property previously in section 5.1.

The structure of graph G can theoretically be anything as long as it captures the conditional independencies in the label sequences being represented. When modeling sequences, however, the simplest and most typical network structure is one in which the nodes corresponding to the elements of Y form a first-order chain. A simple first-order chain is illustrated in Figure 5.3.

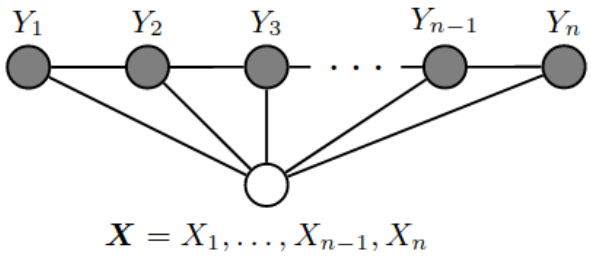

Figure 5.3 : First-order Markov chain

Lafferty et al. define the the probability of a particular label sequence y given input data sequence x to be a normalized product of potential functions, each of the form

$$\exp\left(\sum_j \lambda_j t_j(y_{i-1}, y_i, \boldsymbol{x}, i) + \sum_k \mu_k s_k(y_i, \boldsymbol{x}, i)\right),$$

where t_j is a transition feature function of the entire observation sequence and the labels at positions i and i − 1 in the label sequence, s_k is a state feature function of the label at position i and

the observation sequence, and λ_j and μ_k are parameters to be estimated from training data. The same expression is described architecturally in Figure 5.4. Here Φ is the hidden state parameter function.

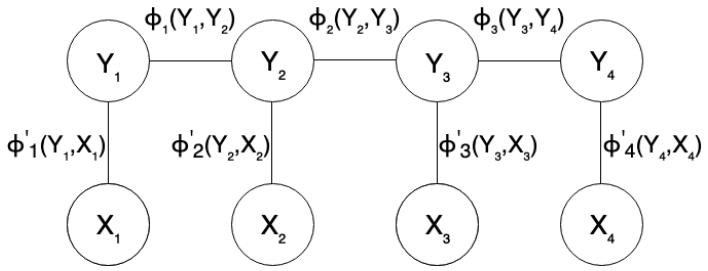

Figure 5.4 : CRF Architecture

5.5.2 CRF For Urdu Word Segmentation

We used a Conditional Random Field model because the Urdu word segmentation problem has now been turned into a sequence labeling task, as described in previous sections. Thus, Urdu sequence labeling is defined as the probability of a sequence of labels Y conditioned over the sequence of input Urdu characters X. The model optimises the value of $\mathbf{P(y_1...y_n \mid x_1...x_n)}$, where $\mathbf{x_1...x_n}$ is a sequence of Urdu characters of a sentence and $\mathbf{y_1...y_n}$ is a sequence of predicted labels. The predicted labels, as described in section 5.2, are part of the L = {B, S, O} set. For ease of computation, the labels were assigned numbers as follows:

- B : Beginning character of a word : **1**
- S : Beginning character of a sub-word : **2**
- O : Other characters : **0**

As an example,

Input String (X) : میںنےراتدن دیکھاہے۔

English Transliteration : *meneraatdindekhaahe*

Predicted Tags (Y) : **['1', '0', '0', '0', '0', '1', '0', '0', '2', '0','1', '0', '0', '0', '0', '1', '0']**

Final Segmentation : مہینے راتدن دیکھا ہے

English Transliteration : *maine raat-din dekhaa he*

English Translation : I have watched day and night

5.6 Results and Analysis

The entire Urdu corpus of 19,651 sentences was split into 17,401 sentences for training and 2250 sentences for testing. We chose F_1 score, precision and recall as our evaluation metrics because they provide an in-depth and reliable assessment of performance, especially in case of an imbalanced dataset. On the unrevealed testing set, our CRF based model attained an F_1 score of 0.92 on sub-word boundary prediction and 0.98 on word boundary prediction. Table 5.8 shows the cumulative improvement in F_1 Score when features are added incrementally. The detailed results and the confusion matrix are shown in Table 5.6 and Table 5.7. Our testing dataset also included randomized samples from the corpus created during previous work done on this same topic and therefore the improved results are compared on a similar scale.

Label	Precision	Recall	F_1 Score
O	0.99	0.99	0.99
B	0.98	0.98	0.98
S	0.9	0.93	0.92

Table 5.6: Precision, Recall and F1 Score corresponding to each label

	O	B	S
O	137294	1126	31
B	1080	47586	29
S	52	169	3008

Table 5.7: Confusion matrix for sequence labeling

Feature	F_1 Score of label B	F_1 Score of label S
4-gram	0.88	0.64
+ Window between tri-gram and 4-gram	0.90	0.82
+ Type of character	0.93	0.84
+ Consonant or Vowel	0.96	0.89
+ Directionality of character	0.98	0.91
+ Frequency of Character	0.98	0.92

Table 5.8: F_1 Score when different features are added incrementally

The results demonstrated by our model outperforms the state-of-the-art methods primarily because of two factors

- Significantly larger data size and richness of the new training corpus.
- The new engineered features which consider various contextual and domain information to give optimal performance.

5.6.1 Result Comparison with Existing Tools

On our testing data set of 2250 sentences, our model achieved an F_1 score of 0.98 for label B (word boundary) and 0.92 for label S (sub-word boundary). On running our model on the dataset used by Zia et al, we achieved an F_1 score of 0.97 for label B and 0.91 on label S. This shows improvement when we compare it with their model's F_1 score of 0.97 and 0.85 for labels B and S respectively.

Label of F_1 score	**Zia et al Model**	**Our Model**
Word boundary F_1 score	0.97	0.97
Sub-word boundary F_1 score	0.85	0.91

Table 5.9 F_1 score comparison of Word and Sub-word boundary prediction

On comparing the performance of our model with existing Urdu tokenizers, particularly Stanford Stanza Urdu tokenizer[44], we see that our model gives the correct word segmentation in many constructs where stanza adds unnecessary spaces as shown in Table 5.10. Stanza Urdu tokenizer, when tested over our corpus gave F_1 scores of 0.82 for label B. The label S was not considered by Stanza in their tokenization pipeline.

Input Urdu Sentence	Stanford Stanza Result	Our Result
کچھ ضرورتمند لوگ Gloss : *kuch zaruratmand log* Translation : Some needy people	کچھ ضرورتمند لوگ	کچھ ضرورت‌مند لوگ
خوشنصیب زندگی Gloss : *khushnaseeb zindagi* Translation : Fortunate life	خوشنصیب زندگی	خوش‌نصیب زندگی
گلت فہمی نہ رکھو Gloss : Galatfehmi na rakho Translation : Don't misunderstand	گلت فہمی نہ رکھو	گلت‌فہمی نہ رکھو

Table 5.10 Segmentation result comparison with Stanford Stanza Urdu tokenizer

5.6.1 Error Analysis

As we reported in the previous sections, F_1 scores of 0.98 and 0.92 for the labels B and S respectively, we have observed and analyzed the scenarios where our model gave a false positive segmentation. The majority of such cases were because of wrong identification of sub-word boundary. A few examples are shown in Table 5.11.

Input Urdu Sentence	Expected Result	Actual Result
مہینے گن کر دیکھا ہے Gloss : *maine ginkar dekha hai* Translation : I have counted	مہینے گنکر دیکھا ہے	مہینے گن‌کر دیکھا ہے

وو مہربان لوگ ہیں Gloss : *wo meherbaan log hai* Translation : They are kind people	وو مہربان لوگ ہیں	وو مہربان لوگ ہیں
کھانا کھا کر سو گئے Gloss : *khaana khakar so gaye* Translation : Slept after having food	کھانا کھاکر سو گئے	کھانا کھا کر سو گئے

Table 5.11 Error Examples

5.7 Conclusion and Future Scope

We showcased a CRF based word segmentation framework for Urdu language, modeled as a sequence labeling task. Our model achieved state-of-the-art performance with an F_1 Score of 0.98 for word boundary prediction and F_1 Score of 0.92 sub-word boundary prediction. We also contributed a benchmark corpus of Urdu sentences, annotated for the task of Word Segmentation.

For determining word and sub-word boundaries, the model mainly relies on manually crafted features coming from domain and linguistic expertise. For this task, we have looked into the usage of deep neural architectures like Bi-LSTM and neural CRF methods. These models showed subpar results because the limited corpus size was not sufficient to automate the feature crafting and hence our manually crafted features proved to be more optimal. Since the deep neural models are data hungry, our corpus can be extended with a larger annotation exercise. The extended annotations can be done in a semi-supervised fashion by utilizing the provided CRF model for tagging a large dataset and then post-editing the annotations for corrections. In a similar way, transformer based architectures can also be explored once the dataset is sufficiently large. On the similar lines of training architectures, a combination of transformer and CRF models can also be explored for further improvement in performance and accuracy.

Chapter 6

Word Segmentation Tool's Application on Practical Use Cases

In the previous chapters we discussed the various diverse challenges that are faced when processing Urdu text. We then proceeded to showcase our approaches to build an accurate system for the Word Segmentation problem with state-of-the-art results. In this chapter we will showcase the application of our word segmentation tool in studying the vocabular differences in Urdu data from different social settings.

6.1 Lexicon from Different Urdu Data Sources

To evaluate and analyze the lexical differences in Urdu data from different social settings, we conducted an extensive study of Urdu data from different sources such as news articles from BBC Urdu, scripts of Urdu soap operas and Urdu tweets and comments from twitter. The processing of this data was done as follows:

1. Web scraping of Urdu textual data from the above mentioned sources.
2. Removal of white space characters across all data points.
3. Applying our in-house Word Segmentation model to tokenize this data into lexical items and words.
4. Listing out the lexical items for each type of dataset.

On the above 3 Urdu datasets of same size from different sources, it was found that, although, there is a sizable overlap in the vocabulary, the lexicon from BBC Urdu is larger in size as compared to that from soap opera scripts and social media interactions. Moreover, we observed that there is an

existence of paired items in the above datasets which are roughly equivalent in their meaning and usage. This was done by picking up random sentences from BBC Urdu source and manually finding their semantically equivalent phrases from the other two datasets. A few instances of this analyses along with their English gloss are shown in Table 6.1

Sentence from BBC Urdu	Sentence with equivalent phrases in Twitter dataset
ان دشوار حالات میں گزارہ **آسان نہیں** in dushwaar halaat me guzara aasaan nahi	**ان مشکل حالات میں** گزارہ **آسان نہیں** in mushkil halaat me guzara aasaan nahi
وہ **چند احباب سے خلوص اور محبت کا ایسا تعلق بھی** رکھتے تھے جس **میں کوئی تکلف نہ تھا** woh chand ahbaab se khuloos aur mohabbat ka aisa talluq bhi rakhtay thay jis mein koi takalouf nah tha	وہ **کچھ دوستوں سے خلوص اور محبت کا ایسا** رشتہ **بھی** رکھتے تھے جس **میں کوئی بناوٹ نہ تھی** woh kuch doston se mohabbat ka aisa rishta bhi rakhtay thay jis mein koi banawat nah thi
یہاں کے **لوگوں میں تعلیم کا تناسب کافی** زیادہ **پایا جاتا ہے** yahan ke logon mein taleem ka tanasub kaafi ziyada paaya jata hai	**یہاں** کے **لوگوں میں تعلیم کا رجحان کافی** زیادہ **پایا جاتا ہے** yahan ke logon mein seekhne ka rujhan kaafi ziyada paaya jata hai
اگر میں آپ کی رعایت سے فائدہ **اٹھا کر بغیر** کرایہ **سامان لے بھی جاؤں تو** میرے **دین**	**اگر میں آپ کی نرمی** سے فائدہ **اٹھا کر بغیر** کرایہ **سامان لے بھی جاؤں تو** میرے **دین** کے **مطابق** یہ

چوری ہو گی agar mein aap ki narmi se faida utha kar baghair kiraya samaan le bhi jau to mere deen ke mutabiq yeh chori ho gi	**کے لحاظ سے یہ چوری ہو گی** agar mein aap ki riayat se faida utha kar baghair kiraya samaan le bhi jau to mere deen ke lehaaz se yeh chori ho gi
اگر دونوں ہستیوں کو ایک کی بجائے **الگ کر دیا** جائے **تو کام** زیادہ ذمہ **داری** سے پورے **ہوں** گے agar dono hastiyon ko aik ki bajaye allag kar diya jaye to kaam ziyada zimma daari se poooray hon ge	**اگر دونوں اداروں کو ایک کی** بجائے علیحدہ **کر دیا** جائے **تو کام** زیادہ ذمہ **داری** سے پورے **ہوں** گے agar dono idaron ko aik ki bajaye alehda kar diya jaye to kaam ziyada zimma daari se poooray hon ge

Table 6.1 Paired phrases with equivalent meanings in different datasets

6.2 Is there a Diglossic Situation in Urdu?

The existence of differences in the vocabulary and the occurrence of equivalent paired words and phrases in different social settings of Urdu raises the question of existence of a Diglossic situation for the language Urdu.

Ferguson first proposed the idea of Diglossia in 1959. According to him, Diglossia is "a relatively stable language situation in which, in addition to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically more complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal

spoken purposes but is not used by any sector of the community for ordinary conversation", Ferguson, Charles F (1959).

The two forms of a language in a Diglossic situation are often referred to as H (high) and L (low). These forms are different from one another on both social and linguistic terms. They differ linguistically in terms of syntax, vocabulary and phonology. While on the other hand, these forms differ socially in terms of literary heritage, prestige and standardization.

Since, vocabular differences are one of the major differences between the two forms of a language in a Diglossic situation, to study these differences it is of utmost importance to have an accurate word segmentation model so that we can get an accurate list of vocabular items from different data sources of Urdu. Hence, our word segmentation model is very crucial in exploring the Diglossic situation in Urdu.

In Urdu language, "Aam Zubaan" refers to the form of everyday or common language that is spoken by the general public, while "Adabi Zubaan" refers to the form which refers to the literary language that is used in literature and formal writing.[45] [46] [47] "Aam Zubaan" is the language that is used in day-to-day conversations, and it can include slang, colloquialisms, and regional dialects. It is the language that is used by the common people in their daily lives. On the other hand, "Adabi Zubaan" is the language that is used in literature, formal writing, and public speeches. It is a more sophisticated and refined form of the language, and it is used by educated people who have a good command of the language. So, the main difference between Aam Zubaan and Adabi Zubaan is that the former is the language of the common people, while the latter is the language of literature and formal communication.

The debate between Aam Zubaan (colloquial language) and Adabi Zubaan (formal or literary language) has been a long-standing topic of discussion among linguists, scholars, and language enthusiasts. Advocates of Aam Zubaan argue that it is the language of the people and reflects the true cultural identity of the speakers, while supporters of Adabi Zubaan argue that it is a refined and sophisticated language that has a rich literary and cultural heritage.

6.3 Conclusion

In this Chapter, we applied our word segmentation model to study the lexicon of Urdu datasets from different social settings. The vocabular differences between these datasets give rise to the question of existence of the phenomena called Diglossia in Urdu. We showed how the social and linguistic situation for Urdu is aligned with Ferguson's theory and the features proposed by him. However, to claim the existence of Diglossia in Urdu, it is required to have a deeper and dedicated study and analyses of Urdu sociology. This initial study opens the doors for the research on Urdu Diglossia and the word segmentation tool, along with numerous applications, can be helpful for the researchers pursuing this topic.

Chapter 7

Conclusion and Future Direction

In this thesis, we have presented our efforts and work done in the research of Urdu text processing. Urdu, being a morphologically rich language, and having no conventional explicit marker for word boundary in the writing system, is a computationally challenging language to process. We have presented our in-depth study and analysis of the various factors which give rise to the word segmentation issues in Urdu writing system. The orthographic and morphological study of the Urdu script gave us directions in understanding and solving the problem at hand. We first applied our learning to build a rule-based system which leverages a lexicon of Urdu words to max-match against. Due to non-contextuality of this approach and other factors discussed, this approach had many limitations. We then decided to tackle the Word Segmentation problem by modeling the task as a Sequence Labeling problem where each character of the script was assigned a label to tell its relative position in a word. We built a standard benchmark corpus for this task by manually annotating Urdu sentences according to the sequence labels, following the annotation rules derived from our study of the Urdu writing system. This corpus was then used to train a sequence labeling model along with feature engineering. Our model has outperformed the existing Urdu tokenization tools and has shown state-of-the-art precision, recall and f1 score. At last we explored and discussed the phenomena of Diglossia for Urdu language in South-Asian countries. We applied our in-house word segmentation tool to study the lexical differences, along with other functional differences, found in the different Diglossic forms of a language.

Albeit the presented tools, approaches, and corpus have produced positive results, a future direction and scope of improvement can still be seen. The model presented in Chapter 5 of this thesis is trained using a manually crafted corpus which is relatively small compared to segmentation benchmark corpora of resource rich languages like Arabic and Chinese. The model also heavily depends on manually engineered features for determining word and sub-word boundaries. A semi-supervised

expansion process of this corpus is one of the natural directions the current work can be extended in order to produce more accurate results. Training the model using latest deep learning architectures like transformer-based architectures can also be explored once a large enough corpus size is achieved. Computationally speaking, the research on the topic of Diglossia and the NLP tools to study and analyze a Diglossic situation are still it their nascent stage. There is a huge space of research and development on the topic of Diglossia waiting on the doorsteps of interested researchers.

Bibliography

[1] https://www.ethnologue.com/language/urd

[2] Durrani, N., & Hussain, S. (2010, June). Urdu word segmentation. In Human Language Technologies:The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 528-536). Association for Computational Linguistics.

[3] Zia, Raza and Athar (2018). Urdu Word Segmentation using Conditional Random Fields(CRFs). Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics.

[4] Afzal, M., & Hussain, S. (2001). Urdu computing standards: development of Urdu Zabta Takhti (UZT) 1.01. InMulti Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International(pp. 216-222). IEEE.

[5] Cai, D., & Zhao, H. (2016). Neural word segmentation learning for Chinese. arXiv preprint arXiv:1606.04300.

[6] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[7] Monroe, W., Green, S., & Manning, C. D. (2014). Word segmentation of informal Arabic with domain adaptation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 206-211).

[8] Rehman, Z., Anwar, W., & Bajwa, U. I. (2011). Challenges in Urdu text tokenization and sentence boundary dis-ambiguation. In Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP) (pp. 40-45).

[9] Ping, G., & Yu-Hang, M. (1994). The adjacent matching algorithm of Chinese automatic word

segmentation and its implementation in the QHFY Chinese-English system. In Proceedings of the 1994 International Conference on Chinese Computing, Singapore (Vol. 301, p. 94).

[10] Wong, P. K., & Chan, C. (1996, August). Chinese word segmentation based on maximum matching and word binding force. In Proceedings of the 16th conference on Computational linguistics-Volume1 (pp. 200-203). Association for Computational Linguistics.

[11] Al-Hejin, B. (2015). Covering Muslim women: Semantic macrostructures in BBC news. Discourse & Communication, 9(1):19–46.

[12] Alexander, M., Dallachy, F., Piao, S., Baron, A., and Rayson, P. (2015). Metaphor, popular science, and semantic tagging: Distant reading with the Historical Thesaurus of English. Digital Scholarship in the Humanities (DSH), 30(suppl_1):i16–i27.

[13] Ali, A. R. and Ijaz, M. (2009). Urdu text classification. In Proceedings of the 7th international conference on frontiers of information technology, (FIT'09), Abbottabad, Pakistan, page 21. ACM.

[14] Alias-I (2008). LingPipe 4.1.0. http://alias-i.com/lingpipe (Last visited: 23-December-2017).

[15] Allan, J. (2012). Topic detection and tracking: Event-based information organization, volume 12. Springer Science & Business Media.

[16] Anwar, W., Wang, X., Li, L., and Wand, X. (2007b). Hidden Markov model based part of speech tagger for Urdu. Information Technology Journal, 6(8):1190–1198.

[17] Anwar, W., Wang, X., Li, L., and Wang, X.-L. (2007a). A statistical based part of speech tagger for Urdu language. In International Conference on Machine Learning and Cybernetics (ICMLC'07), Hong Kong, China, volume 6, pages 3418–3424. IEEE.

[18] Archer, D., Wilson, A., and Rayson, P. (2002). Introduction to the USAS category system. Benedict project report, October 2002.

[19] Artzi, Y., Lee, K., and Zettlemoyer, L. (2015). Broad-coverage CCG semantic parsing with AMR. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, (EMNLP'15), Lisbon, Portugal, pages 1699–1710. Association for Computational Linguistics (ACL).

[20] Azimizadeh, A., Arab, M. M., and Quchani, S. R. (2008). Persian part of speech tagger based on Hidden Markov Model. In Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data (JADT'08), Lyon, France, pages 121–128.

[21] Baker, P., Hardie, A., McEnery, T., Cunningham, H., and Gaizauskas, R. J. (2002). EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Markup and Harmonisation. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02), Canary Islands - Spain, pages 819–825.

[22] Baker, P., Hardie, A., McEnery, T., and Jayaram, B. (2003). Corpus Data for South Asian Language Processing. In Proceedings of the 10th Annual Workshop for South Asian Language Processing (EACL'03), Budapest, Hungary, pages 1–8. European Chapter of the ACL.

[23] Balossi, G. (2014). A corpus linguistic approach to literary language and characterization: Virginia Woolf's The Waves, volume 18. John Benjamins Publishing Company.

[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.

[28] Baudiš, P. (2015). YodaQA: A modular question answering system pipeline. In POSTER 2015-19th International Student Conference on Electrical Engineering, Prague, Czech Republic, pages 1156–1165. Faculty of Electrical Engineering, CTU Prague.

[29] Becker, D. and Riaz, K. (2002). A study in Urdu corpus construction. In Proceedings of the 3rd workshop on Asian language resources and international standardization, COLING 2002 post conference workshop, Taipei, Taiwan, volume 12, pages 46–50. Association for Computational Linguistics (ACL).

[30] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. Journal of machine learning research, 3(Feb):1137–1155.

[31] Bird, S., Klein, E., and Loper, E. (2009). Natural language processing with Python. " O'Reilly Media, Inc.".

[32] Board, U. D. (2008). Urdu Lughat. Urdu Lughat Board, Karachi, Pakistan.

[33] Bögel, T., Butt, M., Hautli, A., and Sulger, S. (2007). Developing a finite-state morphological analyzer for Urdu and Hindi. In Proceedings of the Sixth International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP'07), Potsdam, Germany, pages 86–96. The Linguistics Department, Potsdam University.

[34] Bond, F. and Ogura, K. (2008). Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary. Language Resources and Evaluation, 42(2):127–136.

[35] Bontcheva, K., Tablan, V., Maynard, D., and Cunningham, H. (2004). Evolving GATE to meet new challenges in language engineering. Natural Language Engineering, 10(3-4):349–373.

[36] Bordag, S. (2006). Word Sense Induction: Triplet-Based Clustering and Automatic Evaluation. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06), Trento, Italy, pages 137–144. Association for Computational Linguistics.

[37] M. Bodirsky, M. Kuhlmann, and M. M¨ohl. Well-nested drawings as models of syntactic structure. In Tenth Conference on Formal Grammar and Ninth Meeting on Mathematics of Language, 2009.

[38] T. B¨ogel, M. Butt, and S. Sulger. Urdu ezafe and the morphology-syntax interface. Proceedings of LFG08, 2008.

[39] A. B¨ohmov´a, J. Haji˘c, E. Haji˘cov´a, and B. Hladk´a. The Prague Dependency Treebank: Three-Level Annotation Scenario. In A. Abeill´e, editor, Treebanks: Building and Using Syntactically Annotated Corpora. Kluwer Academic Publishers, 2001.

[40] S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. The TIGER Treebank. In Proceedings of the Workshop on Treebanks and Linguistic Theories, Sozopol, 2002.

[41] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. Computational linguistics, 18(4):467–479, 1992.

[42] M. Butt and T. H. King. The status of case. In Clause structure in South Asian languages, pages 153–198. Springer, 2004.

[43] M. Butt and T. H. King. Urdu ezafe and the morphology-syntax interface. Proceedings of LFG08. CSLI Publications, Stanford, 2008.

[44] Stanford Stanza Urdu Tokenizer https://stanfordnlp.github.io/stanza/

[45] “Urdu Zaban Ki Tashkeel" by Dr. Gopi Chand Narang. This book explores the development and structure of the Urdu language, including its colloquial and literary forms.

[46] "Urdu Adab Ki Tashkeel" by Dr. Saleem Akhtar. This book focuses on the evolution and development of Urdu literature, including its various genres and styles.

[47] "Urdu Zuban Ka Irtiqa" by Dr. Tariq Rehman. This book discusses the historical and sociolinguistic aspects of the Urdu language, including its colloquial and literary dimensions.

Appendix

1. Urdu alphabet with corresponding phonemic sounds