Morphological and Orthographic Challenges in Urdu Language Processing: A Review

Tayyaba Fatima, Raees Ul Islam, Muhammad Waqas Anwar
Department of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan
{tayybafatimah, rtraees}@gmail.com, waqasanwar@ciitlahore.edu.pk

Abstract

Urdu is the national language and lingua franca of Pakistan. It is a grammar enriched language. It has a big variety of derivation and inflections in a single word that makes it a challenging language to work on language processing tasks. Research in Natural Language Processing (NLP) and Computational Linguistics (CL) has composed a considerable measure about the history of Urdu language, evolution of Urdu literature, usage of Urdu language in a wide range, effects of other languages on Urdu, Urdu dialect and script etc. Most of the work done on Urdu in the field of Natural Language Processing (NLP) and Computational Linguistics (CL) is related to its morphology, orthography and script. Urdu has a very rich and complex morphology which makes it a challenging language in Natural Language Processing (NLP) and Computational Linguistics (CL) tasks. The purpose of this paper is to comprehensively review the morphological and orthographic challenges that arise in Urdu language processing. In modern linguistics morphology and orthography has a central place. Other branches like historical linguistic, phonemics and morphonemics are also important. But this work focuses on Urdu morphology and orthography. Few studies highlighting these Morphological and Orthographic challenges in Urdu Language Processing (ULP) can be found in the literature but still there are many unsolved problems that need to be highlighted and solved. This paper is intended to present, group, and review these challenges.

Keywords: Natural language processing, Urdu language, Morphology, Orthography, Resources

1 Introduction

Natural Language Processing has become pretty mature in western languages. But in the case of South Asian languages like Urdu, core linguistic resources e.g. corpora, WordNet, dictionaries, tag sets and associated tools are still not available. Urdu has a complex morphology and orthography. There are very few working Natural Language Processing (NLP) applications of Urdu language available. The main reason is that it is a very difficult language to perform any computational task on it. Another big reason is the limited market and relaxed enforcement of copyright laws in the region, which causes unwillingness of the developers towards developing resources and products for language. While working on Urdu language processing, a lot of challenges arises. Whether the work is being done on speech or it is on text, it becomes a challenge to work on Urdu language processing. Some of these challenges are described in the literature, while many others need to shed light on them. In this paper, we have presented the challenges that are faced while working on Urdu language processing. A detailed analysis has been made to dig out these problems and challenges.

2 Previous Work

Some challenges in Urdu language processing have been discussed in the literature. A comprehensive review of different algorithms and techniques of stemming Urdu text is provided by (Abdul Jabbar and Khan, 2016). They discussed the challenges influencing the stemming process because of morphological richness of Urdu language and inclusion of loan words of languages like Hindi, Arabic, Persian etc. They also display an extensive audit of various algorithms and systems of stemming Urdu text and furthermore, investigated about the language structure, morphological closeness and other basic highlighting and stemming approaches utilized as a part of Urdu-like languages i.e. Arabic and Persian. Moreover, they investigated benefits and weaknesses of the utilized stemming approaches in Urdu text stemming. (Ali Daud, 2016) had provided a comprehensive survey on Urdu language processing. The center target of the study is to provide an overview with respect to various linguistics assets that exits for Urdu language processing. The study endeavors to portray in detail the current advances made in Urdu language processing. At first, the accessible data-sets for Urdu language are discussed. Resource sharing amongst Hindi and Urdu, morphology and orthography of Urdu language are discussed. Parts of the text preprocessing tasks such as stop words expulsion, diacritics evacuation, normalization and stemming are delineated. A survey of language processing tasks for example tokenization, Sentence Boundary Disambiguation (SBD), Part of Speech (POS) labeling, Named Entity Recognition (NER), parsing and advancement of WordNet undertakings are talked about. Moreover, effect of Urdu language processing on application regions for example Information Retrieval (IR), classification and plagiarism detection is researched. At last, open issues furthermore, future bearings for this new and dynamic territory of research are given. The objective of their work is to sort out the Urdu language processing work in a way that it can give a stage to research work in Urdu language processing. (Riaz, 2007) has pointed out challenges in writing Urdu stemmer. (Hussain, 2007) has discussed Computational Linguistics (CL) issues in Pakistan and gave some proposals. The study briefly

describes the present work on computational linguistics in Pakistan, issues in its improvement and a few propositions for quickening the current pace of work in computational modeling of Pakistani Languages. (Sajjad Ahmad, 2011) presented issues pertaining to the development of a rule based stemmer for Urdu language. They talked about some rule based English, Arabic, Persian and Urdu stemmers. Extremely less work has been done on Urdu stemmer due to its complex and rich morphology. Other than its own vocabulary, Urdu is likewise affected by other morphologies for example Arabic, Persian, Hindi, English and so on. The paper also called attention to a few challenges relating to the advancement of a Urdu stemmer. These issues ought to be considered while building up a rule based Urdu stemmer. (Zobia Rehman, 2001) discussed Challenges in Urdu text tokenization and sentence boundary disambiguation. They divided tokenization issues into two categories: Space inclusion issues and Space exclusion issues. (Waqas Anwar, 2006) carried out a survey about automatic Urdu Language Processing (ULP). It contains the initial attempts in the development of resources for Urdu language processing. They also presented different type of Linguistics analysis in their work.

3 Challenges in Urdu Language Processing

Urdu language has always proved to be a big challenge in the field of language processing. In this section, we discuss the morphological and orthographic challenges of Urdu language processing that are still unsolved.

3.1 Morphological Richness

Natural language processing and computational linguistics tasks become harder as the morphology of the language becomes more complex. Same is the case with Urdu language. It is a morphologically complex language. There exist a number of variants in Urdu against a single word (Abdul Jabbar and Khan, 2016). It is a rich language in the case of both inflectional and derivational morphologies. In order to work on Urdu language processing and its related areas, one should be very clear about its morphology and morphological system.

3.2 Lack of Standardization

A significant problem faced by researchers of Urdu language is lack of standardization of language writing rules. Still there are no standards about writing Urdu text, for example space inclusion and exclusion issues (Zobia Rehman, 2001) i.e. where to add space between words, how to write joiners and non joiners, no standards about writing compound words, what should be added between non-joiners and compound words i.e. space, zero width non-joiner, no-break-space, zero-width space or zero-width no-break space. Lack of standardization regarding Urdu text writing creates serious issue while doing language processing tasks on Urdu language. These issues of writing Urdu text can be solved if proper standards are made and imposed. Challenges created by such issues are highlighted in the following section.

3.3 Text Tokenization and Language Modeling

Lack of standardization in writing text creates a lot of difficulties in text tokenization and Language Modeling (LM). It must be decided that how the compound words will be treated in the Language Modeling (LM) i.e. Unigram, bigram or trigram. Urdu compound words consist of two or three meaningful words. It must be decided that how these words will be treated; whole word as a unigram or first word, و (*wao*) and second word as three separate words. It can cause problem while performing text preprocessing tasks like tokenization and stemming. In the following section we have discussed the categories of Urdu compound words in detail.

3.3.1 Compound Words

In Urdu language, when two or more stems combine to make a single meaningful stem, it is called a compound word (مرکب لفظ) (*murakkab-lafZ*). These words are often used to create preciosity and stress in words. Following are the major categories of Urdu Compound words (Zobia Rehman, 2001) (Schmidt, 1999):

1. AB Compound
2. A-o-B Compound
3. A-e-B Compound
4. A-al-B Compound

AB Compound This is a compound in which two stems are attached to make a single stem (Zobia Rehman, 2001). Both of the stems are meaningful but second stem is extra in need. This is called (تابع موضع) (*tabE-moZ-u*). Examples of AB formation are given in Table 1. There is also a situation where first stem is meaningful but second stem does not bear any meaning but still attached to first stem to make a compound. This is called (تابع مہمل) (*tabE-mohmil*). Few examples of such words are given in Table 2.

آس پاس	(*Aas-pas*) (around)
آگے پیچھے	(*Aagy-peecHy*) (back and forth)
صاف ستھرا	(*Saf-sutHra*) (neat and clean)
چوری چھپے	(*cHoree-cHupay*) (sneak)
رونا دھونا	(*rona-dHona*) (melodrama)
کھانا پینا	(*kHana-pina*) (eat and drink)
آنا جانا	(*Ana-jana*) (come and go)
لینا دینا	(*laina-daina*) (give and take)
چلنا پھرنا	(*cHalna-phirna*) (to walk)
اِدھر اُدھر	(*idhar-udhar*) (here and there)

Table 1: Examples of AB Compound 1

روٹی وٹی	(*roTee-woTee*) (bread)
میل جول	(*mail-jol*) (interaction)
میلا کچیلا	(*maila-kucHaila*) (unclean)
کوڑا کرکٹ	(*kuRa-kirkiT*) (garbage)
خالی خولی	(*kHalee-kHulee*) (empty)
جھوٹ موٹ	(*jhoT-moT*) (lie)
پانی وانی	(*panee-wanee*) (water)
گول مٹول	(*mela-kucHela*) (gol-maTol)
سوداسلف	(*soda-salf*) (Grocery)
کالا کلوٹا	(*kala-kaloTa*) (Black)

Table 2: Examples of AB Compound 2

A-o-B Compound The compound word can be in A-o-B formation (Schmidt, 1999) (Zobia Rehman, 2001). Two words are linked together by the morpheme (و) (*wao*) making a single meaningful word. Few examples of A-o-B formation are given in the Table 3. In these examples, the word Wao (و) is called (حرفِ عطف) (*harf-e-ETf*). It links two stems making the compound word. It gives the meaning of (اور) (*aur*) (and) between two stems.

شب و روز	(*sHab-o-roz*) (day and night)
نظم و نسق	(*naZm-o-nasQ*) (discipline)
ملک و قوم	(*mulk-o-Qom*) (country and nation)
لیل و نہار	(*lail-o-nahar*) (night and day)
خیر و شر	(*khair-o-shar*) (good and bad)
وسیع و عریض	(*wasee-o-Areez*) (wide)
غرور و تکبر	(*gharoor-o-takkabur*) (pride)
قرب و جوار	(*qurb-o-jawar*) (near and around)
صبح و شام	(*subh-o-shaam*) (morning and evening)
مال و متاع	(*maal-o-mataa*), (assets)

Table 3: Examples of A-o-B Compound

A-e-B Compound In this formation, two stems are attached together by writing ِ (*Zair*) under the last letter of first stem. This makes them a single semantic unit. Some examples of A-e-B formation are given in Table 4.

خدمتِ خلق	(*kHidmat-E-kHalq*) (social welfare)
صبحِ صادق	(*subh-E-kaazib*) (early morning)
اہلِ بیت	(*ehl-E-bait*) (people of the house)
دلِ مضطر	(*dil-E-muZta*) (worried heart)
خانہِ خدا	(*kHana-E-kHuda*) (God's house)
وقتِ رخصت	(*waQt-E-rukHsat*) (leaving time)
براہِ مہربانی	(*barah-E-meharbani*) (kindly)
راہِ راست	(*raah-E-raast*) (straight path)
افواجِ پاکستان	(*afwaj-e-Pakistan*) (Pakistan forces)
صدرِ مملکت	(*sadr-E-mumlikat*) (The President)

Table 4: Examples of A-e-B Compound

A-al-B Compound Sometimes two stems are joined together to make a single stem by a joining morpheme (ال) (*al*) . These types of compound words are one of the basic structures of the Arabic root. Table 5 shows few examples of A-al-B compounds.

بین الاقوامی	(*bain-al-aQwamee*) (international)
باب المدینہ	(*baab-al-madeena*) (door of Madina)
رد الفساد	(*radd-al-fasaf*) (elimination of discord)
بیت المقدس	(*bait-al-muQaddas*) (holy house)
ضرب المثال	(*Zarb-al-amsaal*) (idioms)
دین الحق	(*deen-al-HaQ*) (true religion)

Table 5: Examples of A-al-B Compound

3.3.2 Joiners and Non-joiners

There are two types of characters in Urdu: 1) Joiners 2) Non-Joiners.

Joiners In Urdu language, some alphabets are connected with their preceding alphabet. These are called Joiners. List of all Urdu Joiners is given here. Each

ب پ ت ٹ ث ج چ ح خ س ش ص ض
ط ظ ع غ ف ک ق گ ل م ن ہ ھ ی

Table 6: Joiner alphabets in Urdu

joiner has three different shapes depending upon its position in word. These three positions may be beginning of word (initial), middle (medial) of word and end of word (final) (Abdul Jabbar and Khan, 2016) (Zobia Rehman, 2001). Three shapes of a joiner (م) (*meem*) are illustrated in Table 7. If a joiner occurs at

Initial	میرا (*mera*) (mine)
Medial	خدمت (*Khidmat*) (service)
Final	آرام (*Aram*) (rest)

Table 7: Shapes of joiners in Urdu

the end of a word, it is necessary to separate it from next word by adding space. If space is omitted, it will join both words with each other and gives wrong meaning. Here we take the example of joiner (گ) (*gaaf*). In the sentence (آگ جلادو) (*aag jala do*) (turn the fire on), each word is separated with a space. But if we write them without space, they join each other making the sentence wrong like shown in Table 8. The word (fire)

آگجلادو (*aagjala do*)

Table 8: Example Sentence

(آگ) (*aag*) ends with a joiner (گ) (*gaaf*) which connects itself to the first letter of the next word, if space is not used between them this makes it a wrong word. So space is required between joiners.

Non-Joiners However, some letters have just one and final shape and they may join their preceding letters but do not connect with letters that are written after them and do not change their shape. These are called non-joiners. List of Non-Joiner alphabets is given here. Non-joiners can be written without adding

اآدڈذرزڑژوۂے

Table 9: Non-Joiner alphabets in Urdu

space between them and it does not damage the word or sentence. For example, in the sentences (کاغذپرتصویربناؤ) (*kagaz par tasweer bnao*) (Make a picture on the paper), words are written without a single space between them. It can be seen that omitting space does not affect the sentence, because in these sentences each word is ending with a non-joiner. The sentences in Table 10 are written without a single space between the words.

میزپراخبارپڑاہے (Newspaper is on the table)
مجھے کھانالادو (Bring me food)
بچے کوکھلونے دو (Give the baby toys)
کاغذپرتصویربناؤ (Draw a picture on paper)
اتوارکے بعدسوموارآتاہے (Monday comes after sunday)

Table 10: Space Standardization

It becomes challenging to tokenize such type of words on the basis or space factor. In English languages each word is separated by space but such type of words in Urdu does not need space to separate them. There is a need of standardization for space inclusion between Joiners and Non-Joiners. A standard could be defined that one should include space after each single word whether it ends on a Joiner, Non-Joiner or it is a compound word. For example the sentence with all the words ending on a non-joiner (کاغذپرتصویربناؤ) (*kagaz par tasweer bnao*) (Make a picture on the paper) and the compound word (شب‌وروز) (*sHab-o-roZ*) (day and night), should be standardized to be written as illustrated in Table 11.

کاغذ(space)پر(space)تصویر(space)بناؤ۔
شب(space)و(space)روز۔

Table 11: Non-Joiner alphabets in Urdu

3.3.3 Connected Stems

Sometimes two stems are joined together by omitting the space between them. Last word of the first stem joins the first word of second stem and both combines to make a single stem. For example two Urdu words (کون‌سا) (*kon-sa*) (which one) can also be written jointly as (کونسا)(konsa), (اس‌کا) (*us ka*) (his/her) can also be written as (اسکا) (*uska*). So it becomes difficult to tokenize such type of words because two words could be written as a single word without any space between them. There should be standardization about such words that all the words must be separated with a single space before and after them. For example it could be specified that writing the words (کون‌سا) like (کونسا) is incorrect and writing (کون‌سا) is correct only. Such type of standards could be defined to solve these issues.

3.4 Challenges in Urdu Stemming

Stemming is the process of extracting the root word from any given word (Riaz, 2007). Stemming is performed on both inflected and derived words. A lot of challenges arise while doing stemming in Urdu language. In this section we have discussed the challenges that are face in developing a stemmer for Urdu language.

3.4.1 Stemming the Infixes

In English language, inflectional of derivational words are created by adding affixes to start or end of the stem e.g. "Unavoidable" is created from the root word '"avoid". It has two affixes attached as "Un" is the prefix and "able" is the suffix. If we remove the prefix and suffix, we can get the root word. But Urdu language has a different case. The biggest problem in Urdu stemming is extracting stem from Infixes. Such words do not have any prefix or suffix attached with them rather they are modified through infixes. Some Urdu stems extracted from the infixes are given in Table 12.

It is observed that words having infixes follow certain patterns. Correct stem could be extracted from such words if patterns for such type of words (having infixes) are made (Abdul Jabbar and Khan, 2016). We have tried to list out all the possible word patterns based on length followed by infixes in Urdu language. Such words can be easily stemmed if rules are made following these patterns. These patterns are given in Table 13.

علوم *Aaloom* (knowledge)	علم *Elm* (knowledge)
وکلاء *wukla* (lawyers)	وکیل *wakeel* (lawyer)
وظائف *waZaif*(scholarships)	وظیفہ *waZeefa* (scholarship)
آداب *Aadab* (manners)	ادب *adab* (manner)
کتب *kutab* (books)	کتاب *kitab* (book)
فقراء *fuQraa* (beggers)	فقر *fqr* (hunger)
اشکال *asHkaal* (shapes)	شکل *sHakal* (shape)
اقوام *aQwam*(nations)	قوم *Qaom* (nation)
فوائد *fawaed* (benifits)	فائدہ *faedah* (benifit)
رفقاء *rufQaa* (partners)	رفیق *rafeeQ* (partner)

Table 12: Example Urdu Infixes

3.4.2 Ambiguous Affixes

Some words fall in the list of Affixes but are actually the part of the stem itself. If the stemmer chop these words considering them as affixes, the resultant word becomes meaningless. These types of words must be

Length 4		
1	فعال	(C_1+C_2+ا+C_4)
2	فعول	(C_1+C_2+و+C_4)
3	فعیل	(C_1+C_2+ی+C_4)
4	فاعل	(C_1+ا+C_3+C_4)
5	تفعل	(ت+C_2+C_3+C_4)
6	فعلا	(C_1+C_2+C_3+ا)
Length 5		
1	افعال	(ا+C_2+ا+C_4+C_5)
2	مفعول	(م+C_2+C_3+و+C_5)
3	مفاعل	(م+C_2+ا+C_4+C_5)
4	تفعیل	(ت+C_2+C_3+ی+C_5)
5	فعالت	(C_1+C_2+ا+C_4+ت)
6	فعولت	(C_1+C_2+و+C_4+ت)
7	فعیلت	(C_1+C_2+ی+C_4+ت)
8	فواعل	(C_1+و+ا+C_4+C_5)
Length 6		
1	افتعال	(ا+C_2+ت+C3+ا+C_6)
2	تفاعیل	(ت+C_2+ا+C4+ی+C_6)

Table 13: Infix Patterns in Urdu

identified as a part of the stem but not an Affix. The stemmer must be able to differentiate between real affixes and those that are part of the stems. There can be ambiguity in prefixes as well as suffixes. For example in the word (shop) (دوکان) *-/dookan/*, it appears that this word contain a suffix *-/an/*(ان), if it is removed then it produced a word -/doo-k/(دوک), which is a wrong word that gives no meaning. Similar is the case with ambiguous prefixes. For example in the word (flight) (پرواز) *-/parwaz/*, it appears that this word contain a prefix (پر) *-/par/*, if it is removed then it produced a word (واز) *-/waAz/*, which is a wrong word that gives no meaning.
Such type of words should be treated as an exceptional case (Abdul Jabbar and Khan, 2016). List of such words should be maintained in order to avoid these ambiguities.

3.4.3 Stemming Errors

While developing a stemmer, it is necessary to keep an eye at the details of what documents are being missed, or what documents are being incorrectly retrieved because of stemming errors. There are two types of Stemming errors over-stemming and Under-stemming.

Over-stemming happens when the algorithm removes too much of the suffix. It gives the words that shouldn't be grouped together by stemming, but are. For example three words (guest) (مہمان) *-/mehman/*, (adventurer) (مہم‌جو) *-/muhim-ju/*, (adventures) (مہمات) *-/muhim-mat/* gives an example of over stemming. These three words will be conflated to a common stem (adventure) (مہم) *-/muhim/*, removing the (ات) *-/aat/*, (جو) *-/ju/* and (ان) *-/an/* respectively, considering them as suffixes. That is a correct result in case of first two words but incorrect for third word because (مہمان) is a stem itself. It feels like it contains the stem (مہم) but actually it does not. Such type of stemming errors should be avoided. Over stemming can be avoided by setting a minimum size of the constraint of the derived stem (Abdul Jabbar and Khan, 2016).

Under-stemming happens when the stemmer leaves the suffix attached to the word. This refers to words that should be grouped together but aren't. For example, from word (anger) (ناراضگی *-/na-raZ-gi/* the stemmer will remove the suffix (ی *-/yey/* and the stem will remain (ناراضگ *-/na-raZ-g/* and that is not a valid stem. The stemmer should remove the suffix (گی) *-/gi/* to get the correct stem that is (angry) (ناراض *-/na-raZ/*. The problem of under stemming can be avoided by using techniques like partial-match algorithms (Abdul Jabbar and Khan, 2016).

3.4.4 Stemming the Plurals

Urdu language has two kinds of plurals: Unaltered Plural and Altered Plural.

Altered Plural (جمع مکسر) *-/jamA-mukassar/* is a form where the original stem and the balance are altered and the stem is changed. In Table 14 we have given some examples of altered plurals. Extracting stems from such kind of plurals is very difficult because whole form of the stem gets changed in it. That makes it difficult to get the stem from it. Such words in Urdu language are based on certain patterns. These words can be successfully stemmed if their patterns are correctly identified.

آداب (*Adaab*) (manners)	ادب (*adab*) (manner)
علماء (*ulmaa*) (scholars)	عالم (*Aalim*) (scholar)
خطوط (*kHatoot*) (letters)	خط (*kHatt*) (letter)
وجوہ (*wajooh*) (reasons)	وجہ (*wajah*) (reason)
اشکال (*shakal*) (shapes)	شکل (*shakal*) (shape)

Table 14: Examples of Altered Plurals

Unaltered Plural (جمع سالم) *-/jamA-saalim/* is a plural form in which the original stem and the balance are not altered. Few examples of unaltered plurals are given in the Table 15. Let's take the word (حاضرین) *-/HaZir-een/*. The stem is (حاضر) *-/HaZir/* that remains unaltered in the plural. Extracting stems from this form of plurals can be much easier. Stem remains unchanged in it and some extra words are attached to it. If this extra word is removed, stem can be extracted.

حاضرین (*HaZreen*) (presenters)	حاضر (*HaZir*) (present)
ثمرات (*samraat*) (results)	ثمر (*smar*) (result)
لڑکیاں (*laRkiyan*) (girls)	لڑکی (*laRki*) (girl)
کھڑکیاں (*kHiRkiyan*) (windows)	کھڑکی (*kHiRki*) (window)
ناظرین (*naZreen*) (viewers)	ناظر (*naZir*) (viewer)

Table 15: Example of unaltered plurals

Broken Plurals Extracting stems from broken plurals is also a challenging task in Urdu. There are certain words in Urdu language that have more than one possible plural of them. Such words are considered to be broken plurals of that word. For instance (rituals) - (رسمیں) (*rasmain*),(رسوم) (*rasoom*) (رسومات) (*rasoomat*) any of these can be used as the plural of the word (ritual) (رسم) (*rasm*). In order to stem such words one has to find out all the possible plurals of a word and to take care of all the possible inflections and derivations that a single word can have.

رسم (*rasm*) (ritual)	(rituals) (رسومات, رسوم, رسمیں)
وعدہ (*waEdah*) (promise)	(promises) (وعود, وعدوں, وعدے)
سجدہ (*sajdah*) (prostrate)	(prostrates) (سجود, سجدوں, سجدے)
وجہ (*wajah*) (reason)	(reasons) (وجوہ, وجوہات)
کتاب (*kitab*) (book)	(books) (کتابیں, کتب, کتابوں)

Table 16: Example of broken plurals

3.5 Word Sense Disambiguation Issues

Word Sense Disambiguation is the process of identifying which sense (meaning) of a word is used in a given sentence, when the word has multiple meanings.

3.5.1 Homonyms

Homonyms are the words having same spelling or pronunciation but different meanings and origins. In Urdu language, Homonyms are called (ذومعنی الفاظ). The Urdu word (بیت) (*bait*) holds two different meanings. It may give the meaning of "House" (گھر) or "Poetry" (شعر). Such type of words is context sensitive. In order to get the meaning of this type of word in a particular sentence, we need to know the context of that word. Meaning of that word depends on the position of that word in the sentence. This is problematic in Word Sense Disambiguation (WSD). Such words cause a number of problems for NLP applications like Machine translation, Text to speech and Information Retrieval. Serious orthographic errors may occur if context is not considered while translating such words. While doing information retrieval, in a word like (میں) machine doesn't know whether the word (میں) (*main*) means "I" or it means "in". This type of words creates ambiguity in the text. Few more examples of Urdu Homonyms are given in the Table 17.

Word	Meaning 1	Meaning 2
عاجز (*Aajiz*)	down to earth	fed up
چشمہ (*cHashmah*)	fountain	glassess
اتفاق (*ittefaQ*)	incidentlly	unity
کان (*kaan*)	ear	mine
ملکہ (*malkah*)	expertise	queen

Table 17: Example of Urdu Homonyms

3.5.2 Homophones

Homophones are defined as: "Two or more words having the same pronunciation but different meanings, origins, or spelling". So homophones are the words that give same sound but are different in writing and meaning. Such words sounds the same but spelled differently. In Urdu language, Homophones are called (متشابہ الفاظ) (*mutasHabah alfaaz*). While working on Speech to Text and vice versa, in Urdu language, homophone becomes big a challenge. Urdu letters that give same sound but are different in use. Examples of Homophones in Urdu are given in the Table 18. Urdu alphabets given in Table 19 posses same sound but are different in use. How to deal with them? When we are doing speech synthesis, it is difficult to differentiate between such types of words.

Word	Homophone
روضہ (*Aajiz*) (shrine)	روزہ (*Aajiz*) (fast)
عاری (*Aajiz*) (fed up)	آری (*Aajiz*) (saw)
آم (*Aajiz*) (mango)	عام (*Aajiz*) (common)
نذر (*Aajiz*) (offer)	نظر (*Aajiz*) (eyesight)
شعر (*Aajiz*) (verse)	شیر (*Aajiz*) (lion)

Table 18: Example of Urdu Homophones

Alphabet	Homophones
ا	ع, آ
س	ث, ص
ح	ہ, ھ
ت	ط
ک	ق
ز	ذ, ظ, ض
ژ	ی

Table 19: Example of Urdu Homophones

3.5.3 Diacritical Marks

There are some special characters in Urdu that lies above or below a letter, called Diacritical Marks (علاماتِ تلفظ). Diacritical marks aren't very common in Urdu writing. These are Zabar Ó, Zair Ọ, Paish Ó, Madaa Ó, Shadd Ó etc. Diacritical marks are not compulsory in Urdu. Mostly they are added only to help in pronunciation. Diacritical marks are not often used. These diacritics change the pronunciation and the meaning of the word and differentiate letters of similar shape with each other. If a didactical mark is added on a letter, it changes whole meaning of the word. For example the word (دور) can be used in two meanings by adding Paish Ó or Zabar on (و) (*wao*). If we add Zabar Ó on (و) (*wao*) it becomes (دَور) (*daor*) which means "Era". If we write Paish Ó on و (*wao*), it becomes (دُور) (*door*), which means "Far away". If diacriticals marks are not used on such words, it becomes difficult to interpret the meaning of that word. In the example below, same word (کھلا) is used in two different sentences in two dif-

ferent meanings by using diacritical marks. The word ((کھلا) with Paish ؤ on (ک) (*kaaf*) gives the meaning of "Opening" and with Zair ؤ on (ک) (*kaaf*) gives the meaning of 'Flowers bloom':

کمرے کا دروازہ کھلا
kamre ka darwaZah kHula
باغ میں ایک پھول کھلا
baagH main aik pHool kHila

In this way, use of diacritical marks can change the meaning of whole sentence. Few more examples are given in the Table 20.

Word 1	Word 2
ہُوا (*huwa*) (happened)	ہَوا (*hawa*) (wind)
دُھن (*dhun*) (passion)	دَھن (*dhan*) (money)
تَیر (*tair*) (to swim)	تیر (*teer*) (arrow)
دُم (*dum*) (tail)	دَم (*dam*) (breath)
سَر (*sar*) (head)	سُر (*sur*) (Tone)

Table 20: Example of Urdu Diacritics

3.6 Translation Issues

Urdu language proves to be a difficult language while translating it to any other language like English. The reason is its complex grammar and rich morphology and orthography. Following section describes the translation issues that are faced while translating Urdu text to English.

3.6.1 Translating Idioms

Idioms become extremely difficult to translate from Urdu to any other language like English. A word-for-word translation of Urdu idiom (ضرب الامثال) (*Zarb-ul-amsaal*) is often nonsense or changes the meaning of whole sentence. If we take an Urdu idiom: (آٹھ آٹھ آنسو رونا) (*AatH AaTh Aansu rona*) if we translate it to English word-for-word, it gives the translation: "Eight tears to cry" and that is totally a wrong translation. This causes some serious orthographic issues. Some example Urdu idioms and their word for word English translation are given in the table below:

دانت کھٹے کرنا	(To sour teeth)
چار چاند لگانا	(Put four silver)
اینٹ سے اینٹ بجانا	(Brick by brick playing)
آ بیل مجھے مار	(Bulls come kill me)
بھینس کے آگے بین بجانا	(Buffalo ahead harp)

Table 21: Example of Urdu Idioms

3.6.2 Non-Equivalent words

While translating from Source Language (ماخذ زبان) (*makHaz zuban*) to Target Language (ہدف زبان) (*hadaf zuban*), Non-Equivalent words are those words that have no alternative in the target language. When we translate from Urdu to another language, if there is no equivalent of a word in the target language, then what translation would be used? Non-Equivalent words results in wrong translations. For example, when translating Urdu to English, a word (چائے پانی) (*chaey paani*) is translated as "Tea and Water." But in Urdu language this compound word is actually used as the money and favors given to someone. Few more examples of the words having no alternate in English language are given in Table 22. Similarly the names given to relationships in Urdu language can never find an equivalent in English language. Few examples of the names of relationships in Urdu language having no alternate in English are given in Table 23.

رم جھم	(*rimjHim*)
ساون	(*sawan*)
مصالحہ	(*maSalah*)
حُقہ	(*huQah*)
دھمال	(*dHamal*)
گویا	(*goya*)

Table 22: Example of Non-Equivalent Words

پھوپھا	پھوپھی
تایا	تائی
خالو	خالہ
چاچا	چاچی
ماموں	ممانی

Table 23: Relationship Names in Urdu

3.6.3 Transliteration Issues

Urdu to Roman transliteration is difficult because there is no standardization on the spellings. While doing transliteration from Urdu to roman, different spellings are used by different people. There is more than one way of writing a particular word of Urdu in roman and all of them can be valid because there is no standardization. For example the word (میں) (*main*) can be translated as "Me" or "Main". Some transliterated words of English are also used in Urdu. Few examples of different spellings of same word in roman are given in Table 24.

ہوں	hun, hon, ho
وہ	woh, wo, vo
مجھے	mujhy, mujhe, mujy
ہم	hum, ham, hm
میں	Main, me, mai

Table 24: Transliteration in Urdu

3.6.4 English Loan Words

Urdu is a language that keeps on evolving with time. Many English words are also included in today's modern Urdu language and are also being used by native speakers of Urdu language. People have converted these English words into Urdu according to their own understanding. For example the English word "Editor" and "Editors" are used in Urdu like (ایڈیٹر) (*ediTar*) and (ایڈیٹرز) (*ediTarz*) respectively. English plurals are also used in Urdu by adding the suffix (وں) (*aon*) with the English word converted to Urdu e.g. the word "Editors" is written as (ایڈیٹروں) (*ediTron*) when converted to Urdu language. Such loan words become difficult while stemming or translating in Urdu language. In Table 25 we have presented few examples English loan words in Urdu.

بینک	bank
پینسل	pencil
کمپیوٹر	computer
پروگرام	program
لائٹ	light
پولیس	police

Table 25: English Loan Words in Urdu

4 Conclusion

Urdu is a grammar enriched language. Many problems arise while performing any Natural Language Processing task on it. Current work briefly presented these problems and challenges. The overall goal of this work is to figure out all the morphological and orthographical challenges faced in Urdu language processing and to present the summary of these challenges. This study can help many new researchers that are trying to develop applications for Urdu language. Efforts are being made to solve the problems. Most of these problems identified in this work can be solved by making proper standards for Urdu language processing and a little more effort on computational matters can solve these problems. Problems faced by tokenization and Sentence Boundary Disambiguation (SBD) can be handled more effectively by using statistical methods instead of using rule based approaches. Issue like space inclusion and exclusion can be solved by standardizing the Urdu text writing in all disciplines. The paper provides a quick review of challenges that a researcher can face while working in Urdu Language Processing (ULP).

References

Abdul Jabbar, S. I. and Khan, U. G. (2016). A survey on urdu and urdu like language stemmers and stemming techniques. *Artificial Intelligence Rev.*

Ali Daud, W. K. (2016). Urdu language processing: a survey. *Artif Intell Rev.*

Hussain, S. (2007). Computational linguistics (cl) in pakistan: Issues and proposals. *Future Directions in Information Access (FDIA).*

Riaz, K. (2007). Challenges in urdu stemming (a progress report). *Future Directions in Information Access (FDIA).*

Sajjad Ahmad, Waqas Anwar, U. I. B. (2011). Challenges in developing a rule based urdu stemmer. *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP).*

Schmidt, R. L. (1999). Urdu: An essential grammar. *Lectures in the Department of East European and Oriental Studies.*

Waqas Anwar, Xuan W., X. L. W. (2006). A survey of automatic urdu language processing. *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics.*

Zobia Rehman, Waqas Anwar, U. I. B. (2001). Challenges in urdu text tokenization and sentence boundary disambiguation. *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP.*