Classifying Urdu Verbs Using Rule Based Approach

Muhammad Waseem1, Qaisar Abbas2, Muhammad Ilyas3, Saad Razzaq4, Sadia Niazi5, Ijaz Asghar6
134Department of Computer Science & IT, University of Sargodha, Sargodha, 40100, Pakistan
3Department of Psychology, University of Sargodha, Sargodha, 40100, Pakistan
4Department of English, University of Sargodha, Sargodha, 40100, Pakistan
1waseem27@gmail.com

ABSTRACT
To make dictionaries complete and to keep their size restricted, there is an approach in the linguistic world to equip these dictionaries with morphological information. This module of morphological information is usually known as a morphological analyzer or morphological classifier, which normally contains the complete possible linguistic information about each word for that particular language and it also describes the rules of derivations from the root of a word and its various inflections, respectively. In this work, a classifier for Urdu verbs (CUV) is proposed which is still a challenging research issue, as Urdu is a language of high inflection and derivation. The available stemmers for Urdu do not provide enough information about inflectional and derivational forms of words. Also, morphological classifiers available for Urdu are not worthy of handling various problems and delivering results that prune errors. In our work, a rule-based CUV is designed which is able to classify 63 forms of Urdu verbs successfully out of 66. Available Urdu language processing tools are very rare compared to other higher inflectional languages such as German, Turkish, etc., which have competitive morphological classifiers. However, the studies related to Urdu verb morphological classification are identified and a comparative study is presented in this article. In short, this work is a positive contribution to the community, and it provides sufficient information with promising results specifically on inflectional and derivational forms of Urdu verbs.

KEYWORDS: Urdu Morphology, Urdu Verbs Classification, Rule Based Model, Verb Analyzer, CUV Algorithm.

1. INTRODUCTION
The main purpose of morphological investigation or analysis is to study the change from surface to the lexical portrayal. On the other hand, the converse of morphological investigation is known as a morphological generation. The fundamental approach to do this is the development of transducers or finite state machines which are implicitly bi-directional and hence can be used for both the mentioned purposes e.g., analysis and generation. This is due to the coding of rules that we have to do for once which then can be used for analysis and also for a generation. Morphological analysis or classification can be performed through probabilistic classifiers like Naïve Bayes, K Nearest Neighbors, Decision Trees, Support Vector Machines, etc., but in our work, we present the rule-based approach (length based or affix stripping) to Urdu verb classification. Morphological studies for various languages are discussed in Section 2 of the literature review. Urdu is an Indo-European language and is gotten predominantly from Arabic, Persian and Turkish dialects [1]. As per a guess, there are around 450 million individuals who communicate in Urdu around the world. In Pakistan, this language has a national status while in India, it is considered an official language. Due to Persian and Arabic inheritance, it uses the same right-to-left Perso-Arabic script for writing text. Urdu has free orderliness and rich in terms of morphology [2] and can obtain words from different dialects promptly. Urdu takes over the majority of the jargon from Arabic [3] and Persian. Hindi has a different script, however, Urdu and Hindi both have the same structures in terms of phonology, morphology and syntax. A few etymologists thought about them as two distinct tongues of one language. The primary contrast between Hindi and Urdu is that Hindi uses Devanagari drought composed from left to right while Urdu uses Perso-Arabic drought composed from right to left in Nastaleeq script. Due to the cursive nature of the Nastaleeq script, the
letters in a word normally connect with each other. Like other scripts, context-sensitivity is present in it and is normally considered to be a complicated system of writing. In Urdu Zabta Taklif 1.01 [4], Urdu consists of 57 letters but according to Siddiqui Urdu has 52 letters\(^1\). The difference of opinion is due to some disputed number of letters in some words and vowels. Urdu character set is divided into three parts, which are alphabets, vowels and other symbols.

Urdu is a morphologically rich language and is able to borrow words from other languages readily. Urdu morphology is a complex phenomenon. Many regional languages besides Arabic, Persian and Hindi have a deep impact on Urdu formation and development. Derivational affixes also regularly occur in Urdu. Words generally have no spaces between them but for computational purposes separated by spaces. Morphological studies in Urdu are quite low in number, however, the studies that have some similarity are discussed in Section 3 of related work. Then we portray aspects of Urdu morphology shortly for the better understanding of a reader in Section 4.

In this work, our focus is on the single-word verb. Inflection of single verb classification is the multiple of tense, aspect, gender, number and mood, summing up to 66 different classified forms which we attempted to generate using rule-based approach presented in Section 5. For representing correct tense, aspect and mood, many verb auxiliaries are also used with these verbs. These verb auxiliaries do inflect with respect to tense, aspect, mood, gender and number. A regular verb is basically represented in three basic stem classifications, which are root, causative one and causative two. Morphology applied to the root form can also be applied to the causative one and causative two forms. The algorithm designed can classify the form of verbs and is presented in Section 5 along with the interface and running example (Section 5.1). Results are critically analyzed in comparison to related work in Section 6.

2. LITERATURE REVIEW

In [5], Koskenniemi's proposed a two-level morphological model. This was the first model in computational linguistics for the investigation of morphologically rich language. However, still, only two well-known strategies are existing yet with respect to our work for the investigation of morphology. One strategy is stemming, and we can observe a number of stemmers exist for languages around the globe with their level of accuracy. This strategy of stemming was spoken to by Porter in [6] and another strategy is morphologically parsing [7]. This method of parsing was first portrayed by Koskenniemi in his work on two-level morphology [5].

One of the two strategies discussed above is Porter's algorithm [6]. It is comprised of two phases. The first phase deduced common endings from inflected words known as de-suffixing, and the second phase added a potential consumption component to the string acquired in the former point. This is known as a re-encoding phase. The two phases could perform successively or simultaneously. The algorithm's distinctive feature was that it didn't utilize a dictionary, which made it suitable for morphological investigation.

Another strategy from above is the Koskenniemi model [5]. It contemplated morphological investigation and morphological change in various situations with respect to phonological information. It used a finite state transducer (FST) to encrypt correspondence. This correspondence was between the root/surface form of the word and the word itself, for example, a word form: fly + s from root/surface form to flies.

KIMMO by Karttunen [8] was the first morphological analysis system that had adopted the Koskenniemi model. This system contained two sections. The first section was about rules and the second section was about lexical features. These lexical features included morpheme construction and morphosyntactic constraints. Techniques of machine learning (ML) were used for obtaining morphological rules discussed in the first section of KIMMO, while supervised and unsupervised methods of ML could usefully classify the two components in the second section of KIMMO. The supervised learning methods [9-10] could exploit the morphological information from the lexical database e.g., CELEX2 was one of those vocabulary databases. On the other hand, unsupervised learning methods [11-13] used a words-list with or without frequency along the way.

3. RELATED WORK

In 2004, Sara Hussain [14] purposed a finite state morphological analyzer for Urdu. It was based on a directed acyclic morphological structure and was named MORPH. Sara analyzed frequently used Urdu verbs first, and then she focused on frequent nouns, and finally studied affixes and presented closed-class words. An online available lexicon\(^2\) was used for comparative analysis. Words with higher frequency (as per use) were chosen for analysis only, hence a limited study. It could not handle all aorabs\(^3\) (diacritics) given in a text file and hence all were ignored during the preprocessing step.


\(^2\) http://www.lc-star.org/docs/LC-STAR_D1.1_v1.3.doc

\(^3\) Aorabs are short symbol which change the sound of word in Arabic and Urdu as well.
In 2006, Humayon et al [15] proposed work on Urdu morphology, orthography and lexicon extraction. Humayon chose Urdu functional morphology 4 because Russian, Italian, Spanish, Swedish and Latin functional morphology was successfully implemented before this work. He described that morphological analysis could be done on a single word or words in combination. In his work of single word analysis, each word in the sentence was analyzed separately. In combination, words with postpositions and auxiliaries were analyzed. This study focused mainly on noun clitics. It further shed light on limited verb-auxiliaries, and finally, composition of adjective clitics was presented.

Tina Boegel, et al [16] purposed a finite state morphological analyzer (FSMA) for Urdu/Hindi. A number of issues were addressed for building an FSMA. Among the issues, potential ambiguity and the non-concatenative morphological issues were particular. Via a cascade of finite-state transducer approaches, the work allowed similar treatment for both Urdu and Hindi languages. Transducer translated both Urdu and Hindi scripts to a common ASCII transcript. The transilator system was based on Xerox finite-state technology tools. No compatibility issue arose in transilator. However, the major issue in designing a transilator system was that both Hindi and Urdu were written in different scripts. This issue was resolved by processing both scripts to a common script.

Studies presented are not exactly according to the line that we have chosen in our work which is affix stripping or length-based approach. These related studies have shown some similarity with respect to encoding of rule and the existing accuracy for the length-based models is 79.63% [31-30].

4. URDU MORPHOLOGY ASPECTS

In morphology, words belong to a particular lexical category, which can be nouns, verbs, adjectives, prepositions, adverbs, etc. Some are discussed as follows.

4.1. Noun Morphology

Urdu nouns are categorized by gender, number, and case. Nouns can be masculine or feminine with respect to gender, for example, laRKA/boy is a masculine noun whereas laRkI/girl is a feminine noun. In case of number, nouns can be singular and plural, for example ghHaRi/Watch is a singular noun and ghHaRiAN/watches is a plural noun. The noun has three cases, which are direct, oblique, and vocative, for example, laRkA/boy is the direct case while laRkI/boy is an oblique case of noun laRkA/boy. The vocative case has two more forms, which are singular and plural. The singular vocative form is laRkI laRkO/boys of noun laRkA/boy and plural vocative case is laRkO/boys of noun laRkA/boy. Nouns may be further divided into two classes based on declension or group, which are called type-I (marked) and type-II (unmarked) nouns. The basic difference between these two categories is that the former has characteristic terminations in the direct singular while the latter does not. Nouns marked for gender are called marked nouns while nouns that have no special gender suffix are called unmarked nouns. Noun ending on masculine or feminine suffix is called marked noun for example laRkA/boy and laRkI/girl are masculine marked nouns whereas laRkI/girl and laRkA/boy are feminine marked nouns. Nouns that do not end on gender suffix are called unmarked nouns, for example kAm/work and kItAb/book are unmarked nouns.

4.2. Pronoun Morphology

There is no gender distinction for Pronouns. Same Pronouns are used for both masculine and feminine. Pronouns are categorized with respect to number as singular and plural in all persons. For example, first-person single pronoun is Ham/we. Gradation of acquaintance or regard is used in second and third persons, for example tu/you and tum/you pronouns are used for acquaintance while aAp/you pronoun is used for regard in second-person. In third-person distance from speakers is also involved, for example ro/one and yAh/this. Pronouns take direct, oblique and vocative cases in the same manner as nouns can take. Nominal pronouns most commonly occur as the undergoer of verbs except for transitive verbs in perfect tenses where subject takes ne/for. Similarly, when a pronoun is trailed by postposition ko/for, ma/men/for, and se/for, it occurs in an oblique case. The second person recognizes three degrees of acquaintance that reveal connection among speaker and individual tended to or speaker disposition towards that individual.

4.3. Adjective Morphology

Adjectives are categorized as marked/stamped and unmarked/plain adjectives. Plain adjectives have an exposed stem in the illustration. Unmarked or plain adjectives have no special suffix and do not change to show agreement, for example, azAh aAm/fresh-mango, etc. On the other hand, marked adjectives have a variety. It can change its inflection with the gender of the following word, for example, akHAm/kAm/good-work, akHAm/bA/t/good-talk, etc. Here adjective aAm/`good and AkHAm/`good are marked for the masculine and feminine property of the following words. Stamped adjectives concur with nouns and have the ability to

4 Functional morphology is the analysis of the mechanical and evolutionary relationship of anatomical form to organism behavior and dynamics.
adjust itself with the gender, number and case with a little arrangement of inflectional post fixes. In contrast, plain adjectives have just single invariant structure. In conclusion, the inflection of adjectives depends on the inflection of nouns [17-18].

4.4. Adverb Morphology

Urdu verbs have two different groups of words, one which are original adverbs called underived adverbs and another is originally other class of verbs but used as adverbs. According to Schmidt [19], Urdu adverbs are divided into five classifications, which are called adverbs of time, place, manner, degree and modal. Adverb of time describes the time where an event is happening, for example, ‘HamESah’/always, ‘kal’/yesterday, etc. Adverb of place describes the location where an event happened, for example, ‘biAhir’/out, ‘andar’/in, etc. Adverb of manner describes how an event happened, for example, ‘UN karo’/do-like-this, ‘UN klEyE/do-like-this, etc. Adverb of degree describes the intensity of happening of event, for example, ‘baRA baRA aAdm’/big-man, ‘ziyAdah ganDa’/more-dirty. Modal adverb [20] modifies a verb or adjective. For example, ‘sirf’/only, ‘SAhld’/Perhaps, etc.

Adverb does not change the form with the sentence structure. Mostly derivational information is learned from the adverb form. It shows what components the word has been shaped. If an adverb appears in a text, then it is searched in the dictionary as it appears in the text. Most of the adjectives and adverbs fall into the following categories which have the concept of proximity like near and far, demonstratives pointing to some object, interrogatives having the concept of questions and relatives.

4.5. Verb Morphology

Urdu verb is more perplexing when contrasted with other word classes [21]. The verb can appear in different forms to show variations in the happening of activities or a particular event. The verb can be classified as intransitive, transitive and di-transitive with respect to a number of object phrases and subject phrases taken by these verbs [22]. Intransitive verbs are those verbs, which take zero verb phrases. Transitive verbs are those, which take at least one verb phrase with respect to the subject phrase. Di-transitive is those, which take at least two or more verb phrases with respect to subject phrases. Verb or predicate can form complex predicates and can have regular or irregular morphology. Transitive and Intransitive verb forms have regular morphology and are called regular verb forms [23]. Di-transitive forms are called irregular forms having irregular morphology. Different morphological classifications that show variation in verb meaning [24] are root, perfective, imperfective, imperative, subjunctive, infinitive, causative one and causative two forms.

Urdu has a rich morphological verb. Verb forms are commonly classified into five groups which are called infinitive, perfective, repeated, subjunctive and imperative. Urdu also has three stem forms, classified as base, one causative and two causative forms. By adding these three basic forms to different morphemes, they are further divided into gender, number, person and mood which establishes 60 classifications of verb in respect of forms. Such 60 classifications of verb don’t have all the characteristics of the verb. With these sixty types, many verb auxiliaries and light verbs are used which present different characteristics of verbs. Combining verb auxiliaries is appropriate at the syntactic level. It is important to combine auxiliaries at the syntactic level of the verb. If the auxiliaries are lumped at the lexical level with verb forms, then the requirement for complex agreement is avoided.

As per Urdu verbs, adding certain morphemes creates various inflected forms or classifications. Different auxiliaries and suffixes are combined with these five forms. With this combination, tense and aspect inflections are formed in a wide scope. Tense is indicated through auxiliary verbs in Urdu. As for as the agreement in verbs is concerned, number and person of the argument structure of the sentence is important which plays a vital role in the inflection of verbs.

In this work our focus is on single-word verb. The inflection of single verb expands in terms of tense, aspect, gender, number and mood. For representing correct tense, aspect and mood, many verb auxiliaries are also used with these verbs. These verb auxiliaries do inflect with respect to tense, aspect, mood, gender and number. A regular verb is basically represented in three basic stem forms, which are root, causative one and causative two. Morphology applied to the root form can also be applied to the causative one and causative two forms.

5. METHODOLOGY

In building Classifier for the Urdu Verbs (CUV), our approach is completely rule based without any addition of training datasets, databases or dictionaries, which were practiced in related work [14-15], and others as discussed in Section 3. We use length-based suffix removal approach for cutting suffixes in descending order. More lengthy suffixes are checked and removed before short length suffixes. By doing this, the correctness of the approach is reasonable. The basic rules finalized are implemented in the design of the CUV algorithm, which describes the procedural model [25] to solve our problem of morphological classification or analysis. The pseudocode of the algorithm is as follows.
CUV ALGORITHM: Inputs a verb and provides its 66 different classifications.

1. //Input a verb for the required information
2. Input ← single word x
3. //Initialization.
4. Base_form ← null
5. Infl ect_form ← null
6. Verb_type ← null
7. Index ← null
8. //Check that input verb is not null.
9. If x ≠ null
10. {
11. // if x be in exception list then its morphological information is picked up from that list.
12. If x is in exception_list V
13. {
14. index ← null
15. Baseform ← V.baseform
16. Infl ectform ← V.infl ectform
17. Verbt ype ← irregular verb
18. }
19. // Check for valid verb inflection with respect to standard rules defined
20. Else If name ends with any of inflection form
21. {
22. Index ← indexOf(infl ectionform)
23. Baseform ← substring (0, index)
24. Infl ectform ← substring(index)
25. Verbt ype ← form out of verb form
26. }
27. Else
28. {
29. // if no verb inflection identified then put the input verb as root.
30. baseform ← name
31. infl ectform ← null
32. verbt ype ← Root Form
33. }
34. // Morphological information of the input word is printed.
35. Output: inputword, rootword, baseform, infl ectform, verbt ype
36. // First check the baseform and then Print all the possible 63 inflection forms of the input word.
37. Output: If baseform ← !null
38. {
39. Print baseform + infinitive form
40. Print baseform + imperfective form
41. Print baseform + Perfective form
42. Print baseform + subjunctive form
43. Print baseform + imperative form
44. Print baseform + subjective form
45. Print baseform + causative one form + Infinitive form
46. Print baseform + causative one form + Imperfective form
47. Print baseform + causative one form + Perfective form
48. Print baseform + causative one form + Subjunctive form
49. Print baseform + causative two form
50. Print baseform + causative two form + Imperfective form
51. Print baseform + causative two form + Perfective form
52. Print baseform + causative two form + subjunctive form
53. }
54. // Do not print anything if baseform is null.
55. Else
56. Output: no word
57. }

Procedural description of the algorithm along with an example is presented next. For the extraction of rules, a list available at CRULP is studied. This list contains 1540 verbs which are collected from the Online Urdu Dictionary. Due to morphological issues that list is pruned to 1299 verbs that we finally used in our proposed work. A detailed discussion is available in Section 6.

5.1 RUNNING EXAMPLE

Evaluation of CUV algorithm is best described by running it on a concrete example. If word \textbf{؟ہے}
‘kHElnA’/play runs on the system CUV then it first checks the word in the exception list from line 11 to 18 in the algorithm. If word is found, then output detail is displayed from that except list. Urdu has some verbs which have different inflections than the root form like ‘gyA’/go has ‘gyA’/went. So, there is an exception list added into the algorithm to handle such cases. If word is not found in the exception list then CUV checks its inflection and finds out its root word, inflection and category of the verb. Next our system displays morphological detail of that word according to line 19 to 26 in the CUV algorithm. After displaying morphological detail of the word, all other possible 63 classifications with detailed morphological information are evaluated by the CUV algorithm as depicted from line 36 to 53. The interface of our purposed system while executing the Urdu verb ‘kHElnA’/play is presented in Figure 1 along with the output screen in Figure 2 shortly due to large view.

Table 1. Failure Verb List

<table>
<thead>
<tr>
<th>جانا</th>
<th>پناهہ</th>
<th>نوانہ</th>
<th>میئنہ</th>
<th>گنا</th>
<th>گنا</th>
<th>گنا</th>
</tr>
</thead>
<tbody>
<tr>
<td>جہان</td>
<td>پناہہ</td>
<td>نوانہ</td>
<td>میئنہ</td>
<td>گنا</td>
<td>گنا</td>
<td>گنا</td>
</tr>
<tr>
<td>فرا</td>
<td>پناہہ</td>
<td>نوانہ</td>
<td>میئنہ</td>
<td>گنا</td>
<td>گنا</td>
<td>گنا</td>
</tr>
<tr>
<td>لیکنہ</td>
<td>پناہہ</td>
<td>نوانہ</td>
<td>میئنہ</td>
<td>گنا</td>
<td>گنا</td>
<td>گنا</td>
</tr>
<tr>
<td>میئنہ</td>
<td>پناہہ</td>
<td>نوانہ</td>
<td>میئنہ</td>
<td>گنا</td>
<td>گنا</td>
<td>گنا</td>
</tr>
</tbody>
</table>

At present our purposed system is limited to verbs only but in future it will be enhanced up to other classes like noun, adjective, pronoun, adverb, etc. Moreover, the size of the Urdu verb list will also be increased accordingly.

Related work discussed in Section 3 is the case of finite state transducers (FST) technology used to study the Urdu verb morphology. In [16][15] and [14], FST based approaches used lexicon heavily to generate morphology and hence not suited to be compared. Further, they didn’t present quantifiable results. Morphological analyzers can be divided majorly in two branches: linguistic and statistical-based approaches. In a linguistic based approach, strategies applied include affix stripping method or length-based method (related to our work), template-based design, and lexicon-based models. Similarly, in statistical approaches, NLP and ML based probabilistic models are devised. The accuracy achieved till to date in length-based approaches is 79.63% [31-30]. Our approach does not rely on lexicon. Rules are hardcoded in the programming language used and the scope of the work is also narrow with limited test data of 1299 words, which becomes the reason that we are getting 80.29% of accuracy. This percentage of accuracy is better in other designing approaches like statistical.

Inflection and irregular forms are the main issues of failure in detecting the verb morphology. At present, we are successful in identifying 63 out of 66 inflection forms completely (depicted in Figure 2) and the goal is still at distant. CUV has the advantage
of speed and accuracy on limited data in comparison to models discussed in [14-16].

7. CONCLUSION

Only few morphological classifiers are available which focus on encoding rules. The finite state transducer is composed of set of rules and lexicon. Finite state transducer could not handle dynamic inflections too. Our work focuses only on single verb and does not handle complex predicates. Our system is totally rule-based and does not depend upon any lexicon. During design issues like over stemming, most famous irregular verbs, irregular singular masculine perfective form and invalid verbs, were solved first. We test CUV CV on verb list, which was designed by CRLUP. Original list is prone to errors. After careful analysis, list is pruned, and more than sufficient results are achieved by testing the updated list on CUV. This hard-coded rule-based approach has the advantage of speed and accuracy on limited data as compared to related models discussed in Section 3. We have further analyzed that this approach could not be successful for large data set until we have intelligent searching mechanism based on some training data along with the use of some statistical model [26-28]. So, a large cushion is available for the verb classification using statistical models.

REFERENCES


