Automatic acquisition of Urdu nouns (along with gender and irregular plurals)

Tafseer Ahmed
DHA Suffa University
Karachi, Pakistan
E-mail: tafseer@dsu.edu.pk

Abstract

The paper describes a set of methods to automatically acquire the Urdu nouns (and its gender) on the basis of inflectional and contextual clues. The algorithms used are a blend of computer's brute force on the corpus and careful design of distinguishing rules on the basis linguistic knowledge. As there are homograph inflections for Urdu nouns, adjectives and verbs, we compare potential inflectional forms with paradigms of inflections in strict order and gives best guess (of part of speech) for the word. We also worked on irregular plurals i.e. the plural forms that are borrowed from Arabic, Persian and English. Evaluation shows that not all the borrowed rules have same productivity in Urdu. The commonly used borrowed plural rules are shown in the result.

Keywords: Urdu, noun, irregular plurals

1. Introduction

Urdu is an Indo-Aryan language spoken in Pakistan, India and some other countries (Grimes & Grimes, 2000). Urdu is a language that has moderate level of lexical resources. We have a wordlist of 5 thousand words[1], a wordlist with POS including nouns (Humayoun, 2006), an Urdu Wordnet including nouns (Zafar et al., 2012) etc. However, these lists are hand-made and/or hand-corrected. Moreover, most of these works are more than five years old.

The amount and variety of digital Urdu text on the internet is increasing day by day, and a researcher can now collect bigger corpus from more domains e.g. sports[2], fiction[3] and blogs[4] beside the newspaper corpus.

The part of speech noun is an open class in which new words can be added. Moreover, different domains usually have different nouns. Hence, we need an automatic system of acquiring nouns from the daily growing potential corpus.

As Urdu is a morphologically rich language, the acquisition of noun has another aspect. Beside the root word, we need to acquire the morphological information related about its forms. This information helps in generating different inflectional forms of the word. For this task, we need to understand regular and irregular morphology of noun and related POS. Following is a brief description of Urdu morphology. For further details see Schmidt (1999).

1.1 Regular Inflections

The inflectional form of Urdu noun depends on number (singular vs. plural), gender (masculine vs. feminine) and (nominative vs. oblique) form. All the Urdu nouns (animate, inanimate or abstract) have a gender. The noun *dil*[5] 'heart' has masculine gender, and the noun *kitAb* 'book' has feminine gender. The abstract noun *piyAr* 'love' has feminine gender. There are different sets of inflection for masculine and feminine nouns as shown in table 1.1.

	Masc		Fem	
	Nom	Obl	Nom	Obl
Singular	*dil*	*Dil*	*kitAb*	*kitAb*
Plural	*Dil*	*dil-ON*	*kitAb-EN*	*kitAb-ON*

Table 1.1: Morphology of Urdu nouns that does not end with gender inflection

Another feature of Urdu noun is its form on the basis of context i.e. the postposition/case-marker following it. If a noun is followed by a postposition or case-marker (e.g. ergative marker *nE* and accusative marker *kO* etc.), its form may change. Compare the forms of plural of *kitAb* 'book' in following sentences.

(1) us=nE kitAb***EN*** paRHIN
3Sg=Erg book.Pl.**Nom** read.Perf
'She read the books.'

(2) us=nE kitAb***ON*** =kO rakkHA
3Sg=Erg book.Pl.**Obl** put.Perf
'She put the books.'

As (2) has plural of *kitAb* followed by *kO*, the oblique form is used. The nominative and oblique forms of Urdu nouns are shown in table 1.1 and 1.2.

There is another kind of Urdu nouns (that are not described in table 1.1). The nominative singular forms of

[1] Urdu 5000 most Frequently Used Words, CRULP, Lahore, 2008. Available: http://www.cle.org.pk/software/ling_resources/UrduHighFreqWords.htm

[2] http://cricnama.com

[3] http://www.aanchal.com.pk

[4] http://blogistan.urdusource.com/

[5] Urdu is written in modified Arabic script. Our processed corpus is in the same script i.e. Unicode text in Arabic script. Here, we write Urdu words in roman transliteration to help the reader in reading the words. The transliteration scheme is described in: http://www.lrec-conf.org/proceedings/lrec2010/pdf/194_Paper.pdf.

some masculine nouns end with *-A* (the masculine morpheme), and nominative singular form of some feminine nouns end with *-I* (the feminine morpheme). The paradigm of inflections for these nouns is given in table 1.2.

	Masc		Fem	
	Nom	Obl	Nom	Obl
Singular	*laRk-A*	*Lark-E*	*laRk-I*	*laRk-I*
Plural	*Lark-E*	*lark-ON*	*laRk-IAN*	*laRk-ION*

Table 1.2: Morphology of Urdu nouns that end with gender inflection

The other POS of Urdu also have inflectional morphemes, and some of these have same forms as of inflectional morphemes of Urdu noun. The morphology of inflecting Urdu adjectives is given in table 1.3.

	Masc		Fem	
	Nom	Obl	Nom	Obl
Singular	*Bur-A*	*Bur-E*	*bur-I*	*bur-I*
Plural	*Bur-E*	*Bur-E*	*bur-I*	*bur-I*

Table 1.3: Morphology of those inflecting Urdu adjectives

In table 1.1-1.3, we observe that the morpheme *-A* is used as the masculine marker for both nominative masculine singular inflected noun and nominative masculine singular adjective. Similar *-ON* is common for uninflected masculine noun, inflected masculine noun and uninflected feminine noun. Similarly, we find homograph[6] of *–A*, *-E*, *-I* and *–ON* etc. in the list of verb inflections given below.

(3) **List of Inflections of Urdu verbs:** -A, -I, -E, -IN, -t-A, -t-I, t-E, -t-IN, -ON, -EN, -O, …….

Hence, we cannot make simple rules like "any word having *–X* morpheme is masculine plural noun, and any word with *–Y* morpheme is feminine singular." We need to work on the whole paradigms of noun, verbs and adjectives (almost all the adverbs and quantifiers etc. follows adjective's morphological paradigm).
For this reason, we worked on matching the whole (or a significant part of the) paradigm in experiment 1 described in the section 2.1.

[6] In Urdu, there are two types of homographs. Some homographs are homophones too. However, there are homographs that are not homophones. The reason is that Urdu (Arabic) script uses diacritic marks. These diacritics are usually not written in Urdu text. Hence, the words or inflections with different sounds can end as homograph. For example, characters ی and یں are used to represent the inflection *-EN* in *kitAb-EN*. The same character sequence is used to represent the inflection *-IN* in *paRH-IN*. (The diacritic *zer* is not written.)

1.2 Irregular Morphology

The tables 1.1-1.2 show the regular morphological rules for Urdu noun. These rules apply on native as well as borrowed words of Urdu. For examples, Arabic *kitAb* or English *computer* also follows the Urdu inflectional rules for plural. However, sometime the plural form from other languages is also borrowed (Schimdt, 1999). For example, we have both *kitab-EN* (the native Urdu rule) and *kutub* (the Arabic plural rule) as the plurals of *kitAb* in Urdu. Similarly, we have *kampiuTer-ON* and *kampiuTer-s* for English word computer. Hence, an analysis of plural paradigms must involve the irregular or "broken" plurals borrowed from other languages. We do not know any empirical study about the percentage of these irregular plurals used in Urdu corpus. Hence, exploring the productivity of different types of irregular plural rules is another goal of this study.

1.3 Context

We are focusing the inflectional rules (specially for plurals) to find Urdu nouns from the corpus. However, we cannot find all the nouns in this way. Most of the mass nouns (e.g. *rEt* 'sand') and many abstract nouns (e.g. *bHarosA* 'trust') do not have plural forms. The singular or plural forms of some nouns may not be present in the corpus due to data sparsity.
For such cases, we can use context information to find nouns and its gender. In Urdu, many modifiers (e.g. ordinals, possessive pronouns, *vAlA*, reflexive pronouns and adjectives etc.) are inflected according to the gender, number and nominative/oblique form of the noun.

(4) pahl-A/accH-A nAvil
first.M.Sg.Nom/good.M.Sg.Nom novel.M.Sg.Nom

(5) pahl-I/accH-I kitAb
first.F.Sg.Nom/good.F.Sg.Nom book.F.Sg.Nom

Hence, working on context (i.e. preceding word) can give a method to find those nouns whose plural forms are not (or rarely) used.

2. Experiments

We conducted three experiments to get the noun lists and other information. The corpus used consists of 8.12 million tokens (Hussain, 2008). There is a wordlist of 50,000 most frequent words compiled from this corpus. The list of word bigrams is also extracted from this corpus.

2.1 Regular Inflections

The first experiment tries to find the noun (and its gender) on the basis of its regular paradigm. The experiment requires filtering the verbs and adjectives that shares some inflections with the nouns. It also requires separating masculine and feminine nouns as there are some common inflectional forms in them.
We made lists of masculine noun inflections, feminine

noun inflections, adjective inflections and verb inflections. The program picked the words from the wordlist one by one. It creates all of its (supposed) inflectional forms in the following order.

First, the verb inflections are appended to word and all of these created inflection forms (potential words) are searched in rest of the wordlist. If sufficient number of these created (potential) words is found in the wordlist, we assign the POS verb to this word, and move to the next word. Otherwise (if sufficient number of verb forms are not matched), then we test for feminine noun, then masculine noun, and lastly for inflected adjective.

These tests must be applied in this strict order. If this order is not followed a word can be found to have more than one POS. For example, *kitAb* is a feminine noun having the forms *kitAb, kitAbEN* and *kitAbON* (cf. table 1.1). If we apply masculine noun test on *kitAb*, the potential words *kitAb* and *kitAbON* will be generated. As these generated words are present in wordlist, one can misclassify *kitAb* as masculine noun.

It shows that we must apply feminine noun test before masculine noun test to get correct guess. Similarly, the order of adjective and noun rules is also important. Consider the example of masculine noun *laRkA*. Its forms are *laRkA*, *laRkE*, *laRkON* (cf. table 1.2). All of these forms are present in the wordlist. If we find the word *laRkA* in the wordlist, and apply adjective inflectional rules on it, then the potential adjectives *laRkA*, *laRkI* and *laRkE* will be generated. As these word are present in the wordlist, one can misclassify *laRkA* as adjective. This problem can be avoided by applying noun test before the adjective test[7].

Hence, the order of the four tests (i.e. verbs, feminine noun, masculine noun, inflecting adjective) ensures the acquisition of correct POS and its feature(s). (If a word passes test for a part speech e.g. say feminine noun, we do not apply further tests e.g. masculine tests on it.)

We search for all the generated forms for nouns and adjectives. However, for verb we considered only those inflections that are not common with nouns and adjectives. Hence, we used 9 verb inflections. We considered the test successful when 5 or more inflected words are matched. The reason is that the number of verb inflections i.e. 9 is comparatively larger than the number of inflections of noun inflections (2 or 3). Due to data sparsity, it is possible that we do not find some of the 9 inflected forms in our corpus/wordlist. So, we choose a number that is larger than the number of noun or adjective inflections, but it is not too large to create chances of missing verb entries in the wordlist.

2.2 Context

An alternate way of finding the noun (with gender) information is to focus on the modifiers of the noun. In section 1, we describe that noun modifiers have inflections agreeing with the noun. Hence, we made lists of inflecting modifiers (genitive marker *kA/kI/kE*, possessive pronouns e.g hamArA/hamArI/hamArE, *vAlA/valI/ValE*, reflexive pronouns *apnA/apnI/apnE*, quantifiers e.g. *itnA/itnI/itnE*, some ordinals e.g. *pahlA/pahlI/pahlE* etc.) for masculine noun and feminine noun. As (inflecting) adjective is an open class, we did not use them in this experiment.

By analyzing the bigrams, the program made two lists. The first list has the words that follow masculine modifiers, and the second list has the words that follow the feminine modifier. Both of these list can have overlapping members e.g. *mErA* (my.masc) *xUbs3Urat* (beautiful) *gHar* (house.masc) vs. *mErI* (my.fem) *xUbs3Urat* (beautiful) *gARI* (car.fem).

The first example suggests that *xUbs3Urat* is a masculine noun as it follows a masculine modifier, however it follows feminine modifier in the second example.

Hence, the raw form of masculine and feminine lists can have adjectives and other "non-noun" words. To filter those words, we find intersection of these two sets. Consider that we term this set of intersecting words as C. Then, two sets are obtained by finding set difference of C from the masculine set (M) and the feminine set (F) i.e. we calculated M-C and F-C. The resulting sets have predicted masculine and feminine nouns, as non-nouns have been removed from these lists (sets). We applied simple regular inflectional rules on each word of resulted lists to convert plural forms in these to their respective root (nominative singular) forms.

2.3 Irregular Plurals

In the experiment described in 2.1, we focused on the regular morphology of Urdu POS including nouns. However, as discussed in section 1, Urdu uses both native plural inflections as well as borrowed plural forms for the plurals of borrowed singular nouns. We termed these borrowed plurals as irregular or broken plural.

In this experiment, we list common plural rules for Arabic, Persian and English. As Arabic is a template based language (McCarthy, 1981) that does not always use affixes for pluralization, we constructed a system that works on templates. We also made a combined list of extracted (singular) nouns from the previous two experiments.

The system picks words from the wordlist one by one. If it is following the template, it generates the corresponding plural form. The generated singular form is searched in the extracted wordlist, and if it is matched, the singular word and the rule number are saved in a list. The rule number is saved to find data about productivity/frequency of each (borrowed) rule.

The system does not tell about the gender. However, the utility developed in section 2.2 helps in predicting the gender of each noun.

[7] The adjective test is irrelevant to our current experiments about noun acquisition. However it will become relevant, if one wants to acquire inflecting adjectives using the same algorithm. Similarly, we do not focus on the result of acquired verb lists, as simple verb is closed class in Urdu, as it uses nominal+verb complex predicate for creation of new verbal predicates.

3. Results

The results of the experiment described in section 2.1 are following.

	Predicted	Correct	Precision
Masc	3540	2667	75.3%
Fem	1175	1063	90.5%

Table 3.1: Prediction of words as nouns on the basis of Inflection

It is observed that the masculine noun list has a lot of misclassified feminine nouns. There are 271 such words i.e. 7.6% feminine nouns in this list. The reason is that the inflection rules for the feminine nouns are applied first, and if the word does not have all of the feminine inflections, then we apply the masculine inflections. As all of the inflections of these misclassified feminine nouns are not present in the corpus (due to data sparsity), they fall to the masculine noun list.

The above fact and table 3.1 shows that the combined precision for nouns without concerning gender is at least 80.4%. (We used the word "at least" because the count of misclassified masculine as feminine noun is not part of this statistic.)

Table 3.2 shows the results of the experiment described in section 2.2. The experiment is designed to get lists of singular nouns. During the evaluation, we found that there were many irregular plurals in the resulting lists. As these plurals do not fulfil our evaluation criteria i.e. singular noun, we do not tag these as correct. However, the results related to these irregular plurals are presented separately in parenthesis.

For example, the experiment gives a list of 2633 predicted masculine singular nouns. The prediction is correct for 1591 nouns that are masculine singular. However, we found that there are 429 plural nouns in this list. We write this number in parenthesis.

Some of these plural forms may collapse to the correctly predicted singular forms, and someone's singular form (i.e. root) may not be present in our predicted list.

	Predicted Singular	Correct (+irregular plurals)	Precision
Masc	2633	1591 (+429)	60.4% (+10.0%)
Fem	1510	1114 (+116)	73.8% (+7.7%)

Table 3.2: Prediction of words as nouns on the basis of context

We find the union and intersection of the evaluated/extracted masculine and feminine noun lists from experiment 2.1 and 2.2. The results are presented in table 3.3.

	All nouns (union of both lists)	Common in both lists
Masc	3274	831 (25.4%)
Fem	1662	514 (30.9%)

Table 3.3: Counts of all (and common) acquired nouns on the basis of inflection and context

Table 3.3 shows that both of these experiments are good for acquiring different sets of nouns.

The results of the experiment about irregular plurals described in 2.3 show that not all plural rules of other languages have same frequency/productivity. The most commonly used borrowed rules are shown in table 3.4.

Affix/Template	Origin	Examples: Singular/Plural	Precision
-z/s	English	*prinTar/prinTarz*	362/406 = 89%
-At	Arabic	*ilzAm/ilzAmAt*	180/252 =71%
$c_1c_2c_3$ -> $ac_1c_2Ac_3$	Arabic	*bdn/abdAn*	48/67 = 71.6%

Table 3.4: Borrowed plural rules of Urdu

The apparent reason for dominance of English plural rule is that a major part of the corpus came from newspapers. If we process a corpus that has literature and/or religious text, we may find different results.

4. Conclusion

We devised three experiments to acquire Urdu noun along with morphological features from corpus. The first experiment focused on inflections of the different forms present in the corpus. We achieve 90.5% precision for automatic acquisition of feminine nouns. The second experiment focuses on the inflection of the preceding word. The precision of its results is less than experiment 1, however we are able to find many nouns that are not discovered in experiment 1. The third experiment focuses on empirically unexplored domain of irregular or "broken" plurals and found that some of the borrowed inflections/plurals are used frequently in Urdu.

In future, the method can be used for acquisition of nouns for other South Asian languages as they have similar rich morphology and similar issues. For relatively less (computational) resourced languages e.g. Sindhi and Punjabi, this method can help in automatic creation of noun lists along with morphological features.

5. References

Grimes J. E. and Grimes B. F. (eds.). (2000). *Ethnologue. Volume 1: Languages of the World; Volume 2: Maps and Indexes*. 14th edition, Dallas: SIL International.

Humayoun M. (2006). Urdu Morphology, Orthography and Lexicon Extraction. Master's Thesis. Chambéry: Chalmers University of Technology.

Hussain S. (2008). Resources for Urdu Language Processing. In *Proceedings of the 6th Workshop on Asian Language Resources, IJCNLP'08*.

McCarthy J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry 12:3*.

Schimdt R. L. (1999). *Urdu: An Essential Grammar*. London: Routledge.

Zafar A., Mahmood A., Abdullah F., Zahid S., Hussain S. and Mustafa A. (2012). Developing Urdu WordNet Using the Merge Approach**.** In *Proceedings of Conference on Language and Technology 2012 (CLT12)*, Lahore, Pakistan.

(Schimdt, 1999)