Computation of Gender of Urdu Nouns

Shair Akbar Khan and Mohammad Abid Khan

Department of Computer Science, University of Peshawar
to_shairakbar@hotmail.com, abid_khan1961@yahoo.com

Abstract

This work is about the computation of gender of nouns in Urdu text which comes under morphology [1]. The system developed for this purpose takes Urdu text as input and processes it for the identification and/or conversion of nouns used in that text, using a system generated database of Urdu nouns. If a noun is not found in the system's database then it is searched in an Urdu Noun dictionary. This dictionary contains information about different nouns. The system then applies the gender conversion rules to the noun if it is an animate one [2, 3, 4], otherwise its context is analyzed for finding its correct gender. This system is designed to help those people who are interested to know the correct gender of Urdu nouns for various purposes like automatic part-of-speech tagging. Further, the system can also be used in Urdu natural language systems with respect to nouns.

1. Introduction

This research paper is about the computational aspect of Urdu Noun Morphology using the gender inflection rules and noun context information observed during the course of this study. The term context used here means the neighboring words of a noun in the current file and its use in other locations in Urdu corpus. The system primarily depends upon a system generated database and on an Urdu noun dictionary, both of which were developed as a prerequisite for the successful implementation of the system. The system is also supported by a set of rules; some of them were identified during the literature study [5].

The system is provided with a user friendly graphical interface, which in turn, is supported by a database. This database is a system generated database. It means that after successful iteration of the system through the dictionary, rule base and context resolution, the system stores the output pair i.e. the name of the noun, its opposite gender and that whether it is a living noun or not. The second important component of the system is an Urdu noun dictionary, which contains the most frequently used nouns along with some of their characteristics. The third part of the system is a rule base. The rule base is constructed to compute the opposite gender of animate nouns. Corpus is another component of the system as shown in Figure1.1. It is a large and structured set of texts usually electronically stored and processed [6]. It can be used to do statistical analysis, checking occurrences or validating the linguistic rules in a specific domain. Here, it is used to find out the context information of nouns used in corpus. The rest of the paper is organized as follows. Section 2 discusses the major system components along with the sources from where the data is acquired and the formulation of rules. Section 3 shows the details of the developed algorithm, its implementation and flow of information inside the system. Section 4 is about evaluation of the software. Section 5 concludes the research paper. The detailed architecture of the system can be understood from Figure1.1:

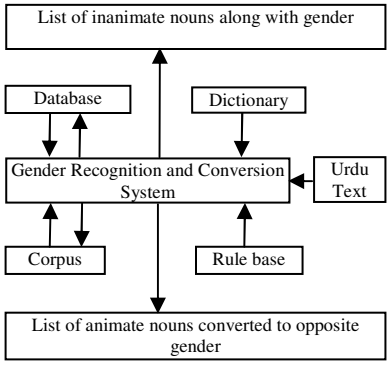

Figure 1.1

2. System components

The following is the detailed discussion of major components of the system as mentioned in Figure 1.1;

namely the database, the dictionary, the corpus and the rule base.

2.1. Database

The purpose of inclusion of database into the system was to provide a source for the fast access of information i.e. whenever the system tries to compute the gender of any noun; it first checks the database for the relevant information before trying the other options. The database was developed using Microsoft SQL Server 2000. The data inside the database was organized in a systematic manner. The database consisted of a table of nouns, which is updated by the system after each successful iteration. The table contains information about nouns i.e. a reference identifier, name, type, gender and the information that whether the noun is animate or inanimate.

The system can be given input in the form of Urdu text; it processes the input file word by word. A word is read from the input file then this word is searched in the database, assuming it as a noun. If the word gets matched with any word inside database, then it is temporarily copied to a file for latter use i.e. to display to the system user. If the word is not found in the database then the system will try other options to compute or to find its gender. Finally, the word will be stored in the database for latter use.

2.2. Dictionary

The online Urdu Dictionary [7] hosted by Center for Research in Urdu Language Processing (CRULP), National University of Computer and Emerging Sciences, Lahore, was accessed for most commonly used nouns and their corresponding morphological information. The required information was downloaded in HTML form, which was then converted into XML. The data in XML format was rearranged locally for efficient access using Microsoft SQL Server. The dictionary consists of a table with the attributes like reference identifier, word, gender and information that whether the noun is animate or inanimate.

2.3. Corpus

The third component of the system is an Urdu corpus, which was organized in a database using Microsoft SQL Server 2000. The corpus normally contains news and current affairs, which was mostly acquired from Urdu websites in Unicode form. This corpus can be used to compute the gender of a noun, if it is not available in the dictionary. The access to the corpus is bi-directional i.e. information can be retrieved from and updated to the corpus, means whenever the system finds new data it will be added to corpus for later use.

If a particular noun is not found in the dictionary then the system uses the context of that noun in corpus to find its corresponding gender. The context information leads the system to decide that whether a noun is masculine or feminine. The corpus can also be accessed by the system for those animate nouns whose conversion to opposite gender cannot be achieved via rule base. Further, a website about Patras Bokhari's [8] work was accessed for data collection, which was used to check the system's performance.

2.4. Rule base

The rule base is the last component of the system. It contains rules for the conversion of animate nouns to their opposite gender. These rules were studied and developed during the literature study. The use of the rule base is done by the system when an animate noun is found in the dictionary. The rules were developed keeping in view the frequency of patterns for conversion. As it is obvious that nouns take a particular method of conversion from one gender to the other, the rules were developed for the most commonly observed patterns. It is worth mentioning here that a noun may incorrectly be converted to its opposite gender by the system. In such a case there is an evaluation process in the system, which was implemented through analysis algorithm. The function of this algorithm is to validate the conversion of nouns, by cross checking the output with corpus. This cross checking process is the inspection of the context of nouns, which are incorrectly converted.

An appendix is given at the end to mention the frequently found patterns in the gender conversion process. Based on the observations given in the appendix, the following rules were defined in rule base. The "0" represents the null character i.e. when there is no gender inflection marker; it is represented by a "0".

Rule No. 1.

If last character of masculine is "ا/ہ/0" then replace "ا/ہ/0" with "ا/ی" to form feminine. This rule was derived from observations 1-4 while converting masculine nouns into feminine ones.

Rule No. 2.

If last character of feminine is "ا/ی" then replace "ا/ی" with "ا/ہ/0" to form masculine. This rule was derived from observations 1-4 while converting feminine nouns into masculine ones.

Rule No. 3.

If last character of the feminine is "ہ" (combined or separated) then remove "ہ" from feminine to form

masculine. This rule was derived from observations 1, 5 and 6.

Rule No. 4.

If the last character of masculine is not "ا/ہ" then add "ہ" (combined or separated) to masculine to form feminine. This rule was derived from observations 1, 5 and 6.

Rule No. 5.

If the last character of feminine is "ن" then replace "ن" with "0/ا /ہ/ی" to form masculine from feminine. This rule was derived from observations 7-10.

Rule No. 6.

If last three or last two characters of feminine are "ئن/ا ئن" then replace "ئن/ا ئن " with "ن/ی" to form masculine from feminine. This rule was derived from observation No. 11.

Rule No. 7.

If last character of masculine is "ن/ی" then replace "ن/ی" with "ا ئن" to form feminine from masculine. This rule is derived from observation No. 11.

Rule No. 8.

If last three characters of feminine are "انی" then remove them to form masculine from feminine. This rule was derived from observation No. 12.

Rule No. 9.

If last two characters of feminine are "نی" then replace them with "0" to form masculine. This rule was derived from observation No. 13.

There is no regular pattern found in observations 14 to 16, therefore the gender of these nouns cannot be computed through a rule. The information about all such nouns was stored in the noun dictionary.

Consider the sentences in Table 2.1, which show masculine and feminine gender of inanimate nouns with the help of context they are used in. Here nouns are used only when they are possessed by someone.

Table 2.1

S. No.	Sentence	Explanation
(1)	یہ میرا قلم ہے- Yih merā qālm hai This my pen is "This is my pen."	The "ا" in word "**میرا**" is pointing to the masculine noun "قلم".
(2)	یہ ہماری کتاب ہے- Yih hamāry kitāb hai This our book is "This is our book."	The "ی" in the word "**ہماری**" is pointing to the feminine noun "کتاب".
(3)	یہ اسکی کتاب ہے- Yih usky kitāb hai This his book is "This is his book."	The genitive case "**کی**" is pointing to the feminine noun "کتاب".
(4)	یہ احمد کا قلم ہے- Yih Ahmad kā qālm hai This Ahmad 's pen is "This is Ahmad's pen."	The genitive case "**کا**" is pointing to the masculine noun "قلم".
(5)	یہ احمد کی کتاب ہے- Yih Ahmad ky kitāb hai This Ahmad 's book is "This is Ahmad's book."	The genitive case "**کی**" is pointing to the feminine noun "کتاب".

The above table contains some gender markers (boldface in the third column) which are very helpful in the gender recognition process. These markers are used by the system to identify the gender of nouns.

There is another issue. Sometimes, a noun's context gives no information about its gender. For example:

(6) کتاب اور قلم لاؤ-
Kitāb aur qālm lāau
Book and pen bring
"Bring book and pen."

In example (6), two nouns "قلم" (pen) and "کتاب" (book) are used where the gender identification is not possible from the context information. To handle such situation, the system was designed to find multiple occurrences of such nouns with varying context information in the corpus. After deciding the gender from corpus, the resultant information is stored in the database. As the last option, if the system is not able to find or calculate the gender of a noun then the expert users will be given an option interactively to store the noun's gender in database.

3. Algorithm development and implementation

The operation of this system is as follows. When Urdu text is loaded, the system will check the end of file. If end of file is true then the program will be ended, otherwise the system will read the next word. The system will decide the word boundary on the basis of blank spaces around the word. As a first step, the system will search the given word in its local system generated database, which is a database of nouns containing their gender information along with their opposite gender if possible. Inanimate nouns have no opposite gender. If the word gets matched in the database, this means that it is a valid noun because the database stores only nouns. The system will then display its relevant information and will loop back to check the end of file. If the word does not match with any other word present in the database then the system will check the dictionary of nouns.

The dictionary contains the frequently used Urdu nouns along with their corresponding attributes i.e. name of noun, type and the information that whether the noun is a living one or not. If the current word gets matched with any other word in the dictionary then it means that the word is a valid noun. After this match, the animate nouns are converted into their opposite gender with the help of rule base. If the conversion is correct i.e. the analysis algorithm returns true then the corresponding noun's relevant information will be displayed and stored in the data base for later use.

The inanimate nouns have no opposite gender, that's why there is no need of conversion. The inanimate nouns are only recognized by the system for their gender. This recognition is done by means of the context of the word as nouns agree for gender to their context in which they are used [9]. The word-context is first checked in the current file. If it doesn't match in the current file then the corpus is searched for finding the gender of the noun. This is the last facility in the system to find the gender of an inanimate noun. Still, if the system fails to find the relevant context information in the corpus then there is a facility provided by the system for expert users to handle such situations. All such nouns can be entered into the database by expert users for later use. The file loading control also provides a facility called tool tip; any word can be selected and clicked for its grammatical information in the file window. The following pseudo code explains the same process:

1-Load Urdu text
2-Read next word from file,
3-If end-of-file then go to Step-10
4-Else look for the word in noun database
  If word is present in noun database then
    Go to Step- 8
  Else look word in dictionary
    If present in dictionary then
      Check noun is animate or inanimate
        If noun is inanimate then go to Step- 6
        End if
    End if
  End if
End if
5-Apply gender conversion rules from rule base and check
If conversion is correct then
  Go to Step- 7
End if
6-Look into the context to find noun's gender
If gender is not found then
Go to Step- 9
End if
7-Store information in noun database
8-Display and Go to Step- 2
9-If user is expert then enter the noun's information to noun database
10-End

The above algorithm was implemented using Microsoft C#, Microsoft SQL Server 2000 and ORM Dot Net [10] on Microsoft Windows XP Professional platform to check its performance and efficiency. Figure 3.1 shows the flow of information within the system:

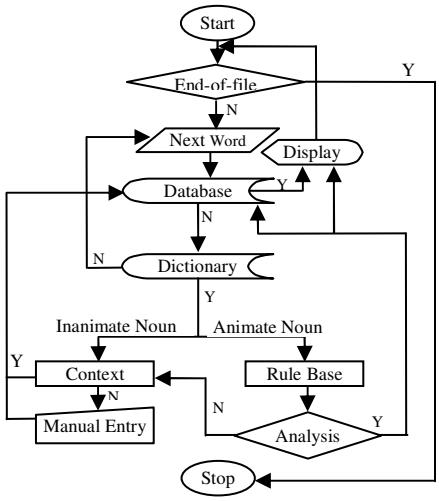

Figure 3.1

4. Evaluation and results

The evaluation of the system was performed by using a corpus of size approximately one giga byte.

The corpus contains data from a verity of domains, but it can be broadly classified under news and current affairs domain. The purpose of the use of this corpus by the system was to evaluate its performance in a given set of data i.e. a file. During the testing of the system, it was observed that 90% of the nouns were present in the dictionary. Later, 73% of the animate nouns were successfully converted into their opposite gender. For the remaining 27% of the animate and for all other inanimate nouns, the system was supported by a corpus. The corpus was accessed by the system for finding the gender of noun and resulted in an overall accuracy of 87%. The remaining 13 % of the inanimate nouns were entered by the system user. After finding a noun's gender through its context, the resultant information was stored into the database. This way the performance of the system can automatically be improved with the passage of time as the system is designed to adopt and store external information.

5. Conclusion

Gender is an important characteristic of nouns in Urdu language. Each noun in Urdu is either of gender masculine or feminine. The aim was to develop an efficient computerized system for the recognition of gender of animate and inanimate nouns in Urdu text and then conversion of animate nouns into their opposite gender. This system is primarily designed to help people who are learning Urdu as a second language. The system will help the learners by deciding the correct gender of nouns. This work can also contribute in the automatic part-of-speech tagging. Further, this work will contribute to the morphological components of Urdu natural language systems.

6. References

[1] W.O. Grady, M. Dobrovolsky and F. Katamba, "Contemporary Linguistics: An Introduction", Addison Wesley Logman, London, 1997.

[2] G. Mustafa, *Jamay ul Qwaed,* Markazi Urdu Board, Lahore, 1973.

[3] M. Abdul-Haq, *Qwaed-e-Urdu*, Anjuman Taraqi-e-Urdu, New Delhi, 1991.

[4] R. L. Schmidt, *Urdu: An Essential Grammar*, Routledge, London, 1999.

[5] S.Hussain, *Finite-State Morphological Analyzer for Urdu,* MS thesis, National University of Computer & Emerging Sciences, Lahore, 2004.

[6] S. Gries and A. Stefanowitsch, "Corpus Linguistics and Linguistic Theory (CLLT)", Mouton de Gruyter, USA and Germany, ISSN 1613-7035, 2005-2008.

[7] Center for Research in Urdu Language Processing (CRULP), Online Urdu Dictionary Service, Pakistan, Retrieved 11- 21- 2007,
Available: http://www.crulp.org/oud/WordIndex.aspx.

[8] S. A. Bokhari, Pakistan Data Management Services, Karachi, 2005, Retrieved 01-02-2008,
Available: http://patrasbokhari.com

[9] P. K. Das, *Grammatical Agreement in Hindi-Urdu and its Major Varieties*, PhD thesis, JNU, New Delhi-67, India, 2005.

[10] Visual Studio.Net, Visual C++, Microsoft SQL Server 2000, Microsoft Windows XP, Microsoft Corporation ®, 2008.

[11] Olero Training Biz Name space, ORM Sample Class Library, (Object Relational Mapping), 2008.

Appendix A

Nouns

The following are the observations during literature study while converting a masculine noun into a feminine one:

Observation No. 1

Delete the last character "ا" of the masculine and add "ی" at the end, to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
اچھا	اچھی	گھوڑا	گھوڑی
اندھا	اندھی	جولاہا	جولاہی
دوھتا	دوھتی	سالا	سالی
کاکا	کاکی	کانا	کانی

Observation No. 2

Add "ی" at the end of masculine to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
میینڈک	میینڈکی	لومڑ	لومڑی
گیدڑ	گیدڑی	کبوتر	کبوتری

Observation No. 3

Replace "ہ" at the end of masculine with "ی" to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
بچہ	بچی	بیچارہ	بیچاری
بندہ	بندی	پگلا	پگلی
شہزادہ	شہزادی	صاحبزادہ	صاحبزادی
کنوارہ	کنواری	نواسہ	نواسی

Observation No. 4
Delete the last character "ا" and add "یا" at the end of end of masculine to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
چوہا	چوہیا	کتا	کتیا
چڑا	چڑیا	گدھا	گدھیا
بوڑھا	بوڑھیا		

Observation No. 5
Concatenate "ہ" at the end of masculine to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
ادیب	ادیبہ	ضعیف	ضعیفہ
بالغ	بالغہ	طالب	طالبہ
ظالم	ظالمہ	عاقل	عاقلہ
فاضل	فاضلہ	قاتل	قاتلہ

Observation No. 6
Add "ہ" (separate) at the end of masculine to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
اداکار	اداکارہ	شاعر	شاعرہ
والد	والدہ	ناصر	ناصرہ
ماہر	ماہرہ	گلوکار	گلوکارہ
قیصر	قیصرہ	عزیز	عزیزہ

Observation No. 7
Add "ن" at the end of masculine to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
چمار	چمارن	سنار	سنارن
رنگریز	رنگریزن	کاگ	کاگن
ناگ	ناگن	مالک	مالکن
لوہار	لوہارن	کمہار	کمہارن

Observation No. 8.
Delete the last character "ی" and add "ن" at the end of masculine to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
باورچی	باورچن	جوگی	جوگن
بلوچی	بلوچن	حاجی	حاجن
بھنگی	بھنگن	پٹواری	پٹوارن
بھکاری	بھکارن	پڑوسی	پڑوسن

Observation No. 9
Delete the last character "ا" and add "ن" at the end of masculine to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
بھٹیارا	بھٹیارن	گوالا	گوالن
دولہا	دولہن	سقا	سقن

Observation No. 10
Delete the last character "ہ" and add "ن" at the end of masculine to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
کنجڑہ	کنجڑن	بنجارہ	بنجارن

Observation No. 11
Delete the last character "ن/ی" and add "ئن/ا ئن" at the end of masculine to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
ترکھان	ترکھائن	قصائی	قصائن
حلوائی	حلوائن	بڑھئ	بڑھائن
چودھری	چودھرائن	مولوی	مولوائن

Observation No. 12
Add "انی" at the end of masculine to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
پنڈت	پنڈتانی	دیور	دیورانی
جیٹھ	جیٹھانی	سید	سیدانی
شیخ	شیخانی	مہتر	مہترانی
نوکر	نوکرانی	چودھری	چودھرانی

Observation No. 13
Add "نی" at the end of masculine to form feminine. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
اونٹ	اونٹنی	مور	مورنی
برھمن	برھمنی	نٹ	نٹنی
بھوت	بھوتنی	بیلدار	بیلدارنی
سور	سورنی	فقیر	فقیرنی

Observation 14 (Exception)
There are certain animate nouns in which no regular pattern is analyzed for gender conversion. Some examples are:

مذکر	مؤنث	مذکر	مؤنث
ابّا	امّاں	پسر	دختر
ابّو	امّی	خالو	خالہ
بادشاہ	ملکہ	خسر	خوشدامن
خاوند	بیوی	خاوند	زوجہ

Observation 15 (Exception)
There are certain animate nouns which are always used as feminine. Some examples are:

آیا	تتلی	قمری	کوئل
بطخ	جنّ	سوکن	کونج
بٹیر	چھپکلی	سہاگن	گلہری
بھڑ	چھچوندر	طوائف	مچھلی
پری	چیل	فاختہ	مرغابی

Observation 16 (Exception)
There are certain animate nouns which are always used as masculine. Some examples are:

کھلاڑی	ٹیچر	بلبل	نالائق
پروفیسر	سیکرٹری	ممبر	یتیم
جانور	صدر	مہمان	میزبان
پرنسپل	رکن	مسافر	وزیر
باز	بگلا	اژدھا	جگنو