Rule-Based Expert System for Urdu Nastaleeq Justification

Muhammad Asad, Ahmed Shiraz Butt, Salahuddin Chaudhry, Dr. Sarmad Hussain
National University of Computer and Emerging Sciences, Lahore, Pakistan
sarmad.hussain@nu.edu.pk

Abstract

Localization of regional languages is a very hot topic these days. Availability of local fonts in this respect is considered a major pre-requisite. This has lead to an information revolution over the globe. Urdu, being a widely spoken language all over the world, needs to be automated in computers and in this regard, a lot of research work is on its way. This paper reflects upon the complexities of justification features in Nastaleeq style of writing Urdu. It further elaborates the problem statement, techniques of justification and recent achievements for its realization in computer in the form of a rule-based expert system.

1. Introduction

Information is central to every society and is considered a basic human right [1, 2]. In recent years, there is a rapid development in ICTs (Internet and communication technologies), which are considered dominant of information. Currently, majority of contents for ICT are in English. This factor substantially limits its effective use in developing countries where very small number of people knows English. Therefore, a concrete solution is required to provide relevant contents in local languages. This has led to the concept of localization of regional languages. To meet the challenges of localization, availability of fonts is a fundamental requirement [3].

On the way to localization, the author with his team worked on a very elegant and much desired feature of Nastaleeq style of writing Urdu language, known as stretching. This feature can ultimately lead to the achievement of justification.

Now the question stands, what is justification? The answer is very simple: To meet the common boundaries for every sentence in Urdu text, irrespective of its varied length of sentences. To achieve this, the Urdu justification methodology exercises two unique techniques namely: stretching and positioning.

Enabling automated justification through stretching and positioning can be very helpful in several real-life scenarios. For example, it can help users of internet so that they would be able to communicate globally without taking pain to manually justify a written text. Designers can produce more creative outputs by using such highly featured fonts. It may also help publishing industry by removing dependencies upon calligraphers for writing title pages and other required hand-written materials. The newspaper industry could significantly utilize this feature where justifying text beautifully in normal contexts as well as in headlines is of prime importance. A lot of manual work could be avoided thus improving the efficiency and saving extra costs.

In this paper, we present the analysis and concept of a rule-based approach to achieve automated justification for Nastaleeq. This rule-based logical model is not only a calligraphic achievement by which the tradition of justification is documented and preserved, but it could also be mapped into an artificially intelligent algorithm, or an expert system, thus facilitating its realization using the tools of computer sciences.

2. Background - the Nastaleeq style

Urdu is the national language of Pakistan, which houses about 145 million people. Globally, it is spoken by over 60 million people in more than 20 countries including Pakistan [4].

Nastaleeq is taken as a standard style for writing Urdu language. It is the most widely used and is defined by well-formed rules passed down through

0-7803-8680-9/04/$20.00 ©2004 IEEE.

generations of calligraphers. Nastaleeq was originally created by the calligrapher Mir Ali Tabrezi, and has been refined by master calligraphers over the past 600 years. Nastaleeq is derived from two other styles of Arabic script Naskh and Taleeq. It was therefore named Naskh-Taleeq, which shortened to Nastaleeq.

2.1. Nastaleeq complexity

Nastaleeq is computationally complex for being highly context-sensitive and cursive in nature. Each letter has precise writing rules relative to the length of the flat nib called *qat*. Shape of a letter depends on multiple neighboring characters [4].

Nastaleeq inherits bi-directional nature of Arabic script. This is particularly true for numbers, which are written from left to right. On the other hand, normal characters are written from right to left [3].

Nastaleeq is a particular writing style in which the letters change their glyph shapes in accordance with neighboring letters. This feature of Nastaleeq is known as context-sensitivity. Normally a certain glyph shape is written in full qat (length of the nib) and is closed before it meets the next joining full qat glyph. The connectors used between glyphs are of at most half qat.

Cursive nature of Nastaleeq is its major defining feature. The letters in a word connect to the next letters in a diagonal manner, from top right to bottom left [3]. This feature poses certain merits and de-merits side by side. Good things include: it takes less width to write a certain word as compared to *Naskh* style of writing Arabic scripts which is primarily horizontal in nature. Limitations include: a word with excessive number of letters in a line may clash with words written in its preceding line.

Other complexities faced are its non-monotonic nature and marks positioning [3]. These complexities, along with other real-life problems, greatly limit its modeling into a workable logical solution for computer implementation. In addition to these, there is a great requirement for spacing and justification issues to be addressed which can make Nastaleeq's computer realization more plausible.

3. Text justification

Every language has different system of features through which it can be distinguished from others e.g. Justification features of Latin script are very much different from Arabic. Justification in English text is achieved mainly by using two techniques: (i) inserting appropriate spaces between words and/or (ii) by increasing/decreasing spaces within the letters of a word. On the other hand Urdu (under Arabic script) justifies a text by stretching and positioning. Though the concept of inter-word spacing exists in Urdu language system, but this is used only as a secondary alternative.

Just like rules defining the methods of writing in Nastaleeq have been passed down through generations of calligraphers, so has been the justification rules. But so far no serious efforts have been made to document and preserve these rules. At this point in time, an Urdu poetic text written using 'Nafees Nastaleeq font' cannot attain any justification, as shown in figure 1. Where figure 2 shows the same poetic sample artistically justified by the calligrapher 'Nafees Shah Al-Hussaini'.

زہے یہ شانِ لاہور، اور یہی شان اس کے شایاں ہے
ہلالِ ضبط و استقلال اُفق پر اب نمایاں ہے

Figure 1: Unjustified computer generated sample

زہے یہ شانِ لاہور، اور یہی شان اس کے شایاں ہے
ہلالِ ضبط و استقلال اُفق پر اب نمایاں ہے

Figure 2: Justified handwritten sample taken from un-published "Hilal-e-Istaqlal"

Figure 2 clearly shows the use of stretching and positioning as tools for justification. Inter-word spacing is however used, but is very insignificant.

3.1. Methodology

Thus, the problem motivated us to figure out the intricacies of stretching, positioning and spacing to attain a justified text.

In our quest to model these features, we sought to learn its underlying principles and calligrapher's intuition. First of all, we took series of regular lectures on Nastaleeq to understand its basic structure, properties, features and limitations. Secondly, issues related to justification were daily discussed in lengthy sessions with the calligrapher. We also grasped the common variations found within calligraphic literature of different calligraphers. Lastly, we opted for Lahore style of writing Nastaleeq by 'Nafees Shah'. We analyzed his several handwritten samples of justified text and other manuscripts and finally succeeded in producing its logical model, as we understood it.

4. Justification in Nastaleeq

The key idea of our approach to achieve justification is not to use spacing as a major technique. Rather significant use of stretching and positioning not only justify the text but also add up to the beauty of the final output.

We now briefly discuss stretching, positioning and spacing to mature our perceptions regarding these techniques.

4.1. Stretching

Using the technique of stretching, a letter of a standard shape and length is replaced with its longer version. This way, a word taking less width takes more space, which ultimately contributes into a longer sentence to meet the common boundary.

The most common tool for stretching is called '*kashish*' and '*madd*'. *Kashish* starts from the first line of grid and goes on till the bottom of 2nd line. It is curved and is similar to a diagonal sword. *Madd* remains on third line and retains the same level. Its shape is like a boat.

Figure 3: Nastaleeq tablet showing kashish of □ and stretched *madd* of □

A certain set of alphabets in a sentence could be stretched in their initial, medial and final positions, whereas very few letters could be stretched in their isolated forms.

4.1.1. Stretched isolated forms. Given are some letters in their isolated normal and stretched forms.

(a) (b)

Figure 4: (a) ں written in normal form, (b) ں written in stretched form

4.1.2. Stretched two character forms. Given are some examples of two character non-sense combinations. Above combinations are in their non-stretched forms while lower are same combinations with initial letters in their stretched forms.

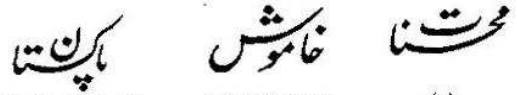
(a) (b) (c)

Figure 5: (a) Stretched ں followed by normal د, (b) Stretched ط followed by لا, (c) Stretched ع medial followed by final

4.2. Positioning

Often, sentences carrying letters of standard length cross the limits of common boundary and shift to the next line. In such situations, stretching could not be used and positioning comes to play its role. Thus, letters in the sentence are overlapped and positioned on one another to adjust the occupied space, enabling the whole sentence to retain within the common boundaries.

Positioning is, without the doubt, the most difficult and complex feature of Nastaleeq and has thousands of variations and rules. Some of positioning examples are:



(a) "Pakistan" (b) "Silent" (c)

Figure 6: (a) 'ن' placed over stretched 'لل', (b) 'ش' is overlapped over 'مو', (c) 'ت' is placed over stretched 'ل' medial

4.3. Spacing

Spacing is one of the less common features of justification in Nastaleeq and is only used in special cases. By definition, spacing has the same meaning as it has for English language system i.e. justifying a text by inserting inter-word and intra-word spaces.

Intra-word spacing is not allowed in Urdu, as it will break a word into its separate constituent letters. Spacing is used in extreme cases, in which a sentence needs to be justified to a bigger length area and either stretchable letters are already stretched, or there exists no (more) letter (s) in the sentence that can be further

stretched and applying positioning results in inappropriate output.

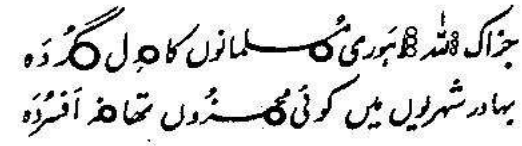


Figure 7: Poetic sample-utilized space highlighted in circles, taken from "Hilal-e-Istaqlal"

Figure 7 clearly shows an extensive use of stretching technique. Here, inter-word spacing appears to be a secondary alternative as no further stretching in the verses is suitable.

It should be noted that from this point onwards, we would limit our discussion to the analysis and results of stretching only. We will discuss positioning only when required and will not state any of its details.

5. Stretching in Nastaleeq

5.1. Classification

Not every Urdu alphabet is stretchable. Some alphabets are never stretched e.g., 'ا', 'و', 'ج' etc. Also, an alphabet stretchable in its final or medial positions may not be stretchable in its initial positions. For example, 'ب' can be stretched in its medial and final positions but it is never stretched in its initial positions.

5.1.1. Letters in initial position. More than 40% of the justification rules apply to stretchable letters when they are in initial positions. Stretching is applicable only if a particular set of letters follow the initial position e.g. س has a longer version but it can only be applied if its next glyph is ا, د, ر, ل, س, م and ه. For all other letters, this stretched version of seen cannot be applied in initial position.

5.1.2. Letters in medial position. The most difficult and interesting part of analysis is the study of letters in the medial position. About 50% of the justification rules apply to stretchable letters in medial position.

5.1.3. Letters in final position. These involve the major use of stretchable *madd* in final position and about 5% of the justification rules apply to this position. Examples include ب, ک, ف etc. However, uses *kashish* in final position for stretching.

5.1.4. Letters in isolation. Isolated letters include those alphabets that can exist in isolation. These are also called *mufrid* in calligraphic terminology. Stretchable *mufrids* are ب, پ, ت, ٹ, ث, س, ش, ف, ک, گ and ے.

5.2. Priority scheme

Calligraphers have always been praised for their artistic ability to utilize the techniques of justification. Just applying the rules of automated justification does not end up in a desired solution. The beauty and richness of resultant justified text that brings comfort to the eyes and give a sense of satisfaction to the native reader is thus the ultimate goal.

To realize this property means developing a *knowledge base* that can simulate calligraphers' tacit ability. This required extracting justification rules and the contexts in which stretchable words have priority over others.

We give an example to illustrate different visual variations of a word that it could possibly assume.

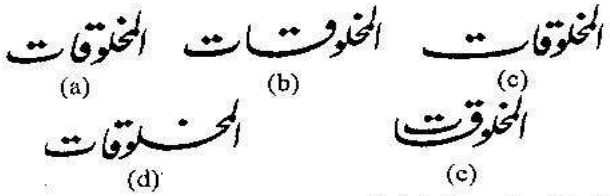


Figure 8: The word "creatures", (a) Non-stretched form, (b) to (e) Stretched forms

In the figure above, all five shapes represent a single Urdu word meaning 'creatures'. The following notations best explain the variations from (a):

(b) ق → stretched ق ا

(c) ث → ثـ

(d) خ → stretched خ ا ل ـ م

From (a) to (e), the first word is written in normal (non-stretched) form. The second and third stretching are the most common practices and even better than

the fourth variation in which خ medial is stretched. Whereas, in fifth shape, both stretching and positioning techniques are used. This form is ideally used and gets the most applause from the native reader. Therefore, such examples motivate us to figure out the best stretchable alternatives among several possible variations, and quantification of their priorities over one another.

Let's look at another example shown in Figure 9:

(a) (b)

Figure 9: (a) □ is stretched, (b) is stretched

The contextual variations are:

(a) ث -> ـبـ | ـ medial

(b) ص -> stretched ـصـ | ـی ـم

It is considered better to stretch ب group (including ب , پ , ٹ , ٹ , ث , ف , ک , گ) on priority rather than to stretch any medial characters preceding it.

5.3. Letter joining constraints

A glyph written using full *qat* (length of nib) never joins another full *qat* glyph; rather it always joins with a half *qat* glyph. But there is an exception in which a full *qat* glyph joins a relatively full *qat* glyph.

(a) (b) (c)

Figure 10: (a) Full qat kashish joined with half qat connector, (b) Full qat kashish joined with full qat circle, (c) Full qat kashish joined with final meem

In the figure above, the word in (a) is correctly written because the stretched ث connects with half qat ی connector in a permissible manner, whereas in (b) the full qat ی connects with full qat ک circle, which is not allowed. The word in (c) is allowed according to calligraphic rules and carries a very sophisticated justification for its validity. It says, "With a very careful observation, one can see that the kashish ending at full qat never joins with a full *qat* ـ at its maximum. Rather, some part of kashish and some of the ـ is already out of contact with each other from upper and lower ends. Thus, kashish is joined with a half *qat* ـ, and ـ is joined with a half *qat* *kashish* of initial ـ"

5.4. Pattern constraints

A poetry text normally follows a specific pattern in which visual representation of certain set of words in each phrase is repeated. A calligrapher tends to justify a text such that some words in each phrase occur at the same point as in other phrases to create homogenous and symmetric effect. Example is given in Figure 11 below.

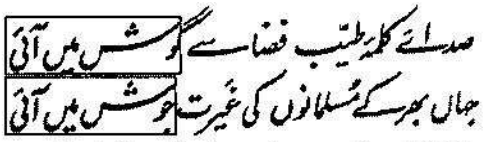


Figure 11: A justified poetic sample exhibiting a visual symmetry (in the boxes)

5.5. Stretching limitation

In a word that has many stretchable letters, there cannot be more than one kashish or more than one stretched *madd*. In simple words, a word cannot have more than one stretched letter. In cases of more than one stretchable possibility, the ties are broken by choosing the most commonly used stretched form. The reason for this limitation is very obvious: the word would seem very lengthy and hence, very un-natural and irritating to a native reader of the Urdu language.

(a) (b)

Figure 12: (a) Only one stretched letter seen, (b) Two stretched letters, seen and noon

In the (b) case where there were two stretchable letters; priority was given to seen because in real-life seen has highest tendency to be stretched.

5.6. Exceptions

Whenever a prose text is selected to be justified, the justification is applied to all the lines except the last line; no stretching is applied in the last line. This is also true for paragraphs where last line of paragraph is not justified.

On the other hand, poetic text is always justified from the very first line till the last one.

6. Implementation results

We developed a rule-based Logical Model, which caters majority of justification issues and present solutions for them. After this step, we moved onwards to implement our Logical Model. The scheme was to develop an application using C# language on .NET platform, which maps the rule-based model into an expert system. Our application has approximately rules.

Following are some examples of justified text that we obtained from our application as results:

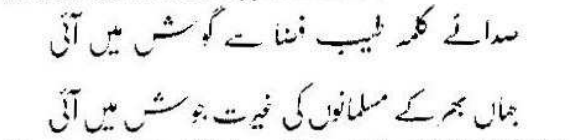
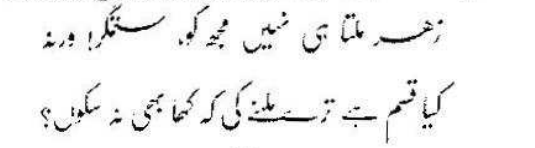


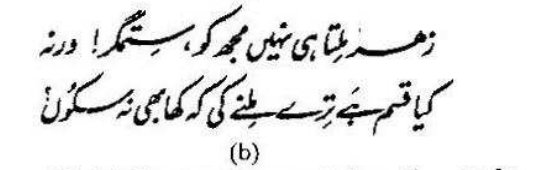
Figure 13: A justified poetic sample obtained from expert system application

7. Discussion

Following are interesting comparative examples:

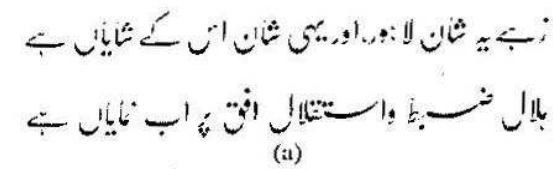


(a)

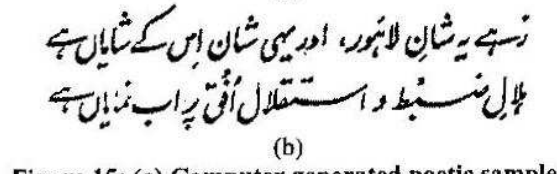


(b)

Figure 14: (a) Computer generated poetic sample, (b) Hand-written poetic sample taken from Diwan-e-Ghalib" by Nafees Al-Hussaini, page 84



(a)



(b)

Figure 15: (a) Computer generated poetic sample, (b) Un-published hand-written poetic sample taken from "Hilal-e-Istaqlal" by Nafees-al Hussaini

In the figures above, the reader can appreciate the reasonable approximation of computer-generated output of justified text in comparison to the handwritten justified text.

However, there is always a room for improvement. There are a number of issues that needs to be addressed. Future works include: horizontal (kerning) and vertical positioning, fine-tuning of existing logical model, and improvement in quality of Nastaleeq font. In our opinion, a dedicated effort is required to make a font exclusively for justification purposes.

8. Acknowledgement

We are very much thankful to our calligrapher *Syed Jamil-ur-Rehman* for writing the glyphs.

9. References

[1] "Info-structure in developing countries," UNESCO's contribution for the World Summit on the Information Society (WSIS) (8 February 2002)

[2] "Statement Presented by Minister of Information and Communications Technology of Thailand," Regional Conference for WSIS in Tokyo Jan 12-15, 2002)

[3] S. Hussain, "www.LICT4D.aisa/Fonts/Nafees Nastalique", in *Proceedings of 12th AMIC Annual Conference on E-Worlds*, Asian Media Information Center, Singapore, 2003.

[4] www.apdip.net/ictrnd/nafees.asp