PrePer: A Pre-processor for Persian

Mojgan Seraji

Uppsala University, Department of Linguistics and Philology
mojgan.seraji@lingfil.uu.se

Abstract

Today web pages through World Wide Web are widely in use as rich resources for developing corpora. These useful source materials contain all sort of texts, including various encodings, and are written by many different authors in various styles. The existence of these factors make Persian text processing complex, therefore, when dealing with Persian, before any natural language processing takes place the input texts need to be prepared and cleaned up into standard texts. A standard text is a text written in standard style where the internal word boundaries are marked based on the official orthography and the style introduced by Academy of Persian Language and Literature (APLL). In the following Sections we will have an overview of text processing issues as well as our solution to pre-process Persian texts by introducing PrePer; a pre-processor for Persian.

1. Introduction

Persian belongs to the Indo-Iranian, a subfamily of Indo-European languages. The language writing system has been greatly influenced by Arabic and has the same characters including four additional letters. As Persian has cursive script, characters have various forms depending on their position in the word. Hence, the characters are differed in orthographic forms and based on how they can be connected to other characters, are divided into two groups: "dual joining" and "right joining" (Seraji et al., 2012). In dual-joining, characters have two distinct shapes depending on their position in the word: initial or medial, and final or isolated respectively. However, three characters in this group, namely ع /'eyn/, غ /qeyn/, and ه /he/ (he-ye do-češm) appear in four distinct shapes. There are also two characters in this group, ط /tâ/ and ظ /zâ/, which have only one shape regardless of their position in the word. However, the right-joining characters do not accept any connection from their left hand side and have only one shape without any distinctive initial, medial, final, or isolated forms.

In computer text representation, there are various sizes and styles of spaces with different Unicode characters, such as no-break space (U+00A0), zero-width non-joiner (U+200B), word-joiner (U+2060), ideographic space (U+3000), zero width no-break space (U+FEFF), and so forth. The use of various space characters of specific width depends on the language characteristics. White space in Persian, designates word boundary as is in many languages. However, there is also another space in Persian, the so-called zero-width non-joiner (ZWNJ, known as, pseudo-space, zero-space, or virtual space) as a boundary inside the word. The ZWNJ is a non-printing character in computerized typesetting placed between two characters to be printed in the final and initial forms to each other. The ZWNJ keeps the word forms intact and close together without being attached to each other.

In Persian, there are various writing styles where the usage of white space might be optional. In official texts such as texts used in mass media, bound morphemes are normally typed with ZWNJ to their adjacent words, while in non-standard language such as in blogs, and forums, these morphemes are usually typed either with intervening white space or in attached form.

Compound words which usually consist of words representing separate lexical categories, may also be typed in different ways. When ZWNJ is ignored and white space is used, words are treated as separated tokens and that can cause problems in tokenizing texts. For example, "حساسیت‌زا" (allergen) may appear either by joining the noun "حساسیت" (sensitivity) to the verbal stem "زا" (to bear, to produce) building one single token (attached form), "حساسیتزا", or as two single tokens delimited, either by space "حساسیت زا" or by ZWNJ "حساسیت‌زا". The optionality of writing compound words as attached single word, detached single words (delimited by space), or as one distinct word (delimited by ZWNJ), in Persian writing system, raises issues in Persian text processing because the frequency of such words will be distributed between different writing styles.

In Persian, inflectional affixes (as pronominal or verbal clitics) also might be written in different forms when it concerns various writing styles; either as attached, detached or with ZWNJ to the adjacent word.

Although Persian and Arabic share almost the same character encoding for the similar letters, there exist a few stylistic disparities when it comes to script such as in the letters "ی" (ye) and "ک" (kaf). These letters can be represented by "U064A" for "ی" and "U06A9" for "ک" in Persian Unicode system and "U0649" for "ي" (Arabic ye has two dots below) and "U0643" for "ك" in Arabic Unicode encoding. Table 1 shows different shapes of the same alphabet for Persian and Arabic.

Due to the fact that different operating systems have traditionally Arabic Unicode characters as their default for many Middle East languages such as Arabic, Persian,

Persian	Arabic	Name of the letters
ک	ك	kaf
ی	ي	ye
ت	ة	te
–	ء	hamza in Arabic
ەی	ەء	he ye (the occurances of ezafe morpheme ye in Persian and hamza in Arabic after the silent "h" /h/)

Table 1: Different forms of Persian and Arabic characters.

Kurdish, Sindhi, Uighur, and Urdu (Saadi, 2007) including codes for additional Persian letters to be used in harmony with Persian, many Persian web pages have texts encoded using Arabic letters as well as Arabic-Indic digits instead of using Persian letters and Persian digits.

Unicode Standard has a separate set of ten characters for Persian digits called Extended Arabic-Indic Digits (Esfahbod, 2004) since three of the ten digits in Persian (number 4, 5, and 6) differ from their Arabic counterpart. Despite existing Unicode characters provided for Persian there are still software that implement Western digits on Persian Keyboards since they do not interpret Persian digits as numerical data (Esfahbod, 2004) and as a result, we see many Persian websites with Western digits entered everywhere. As an example, "Ettelaat.com" which is one of the oldest newspaper in Iran with 87 years of continuous presence in the information, still has a mixed character encoding of Arabic letters and English digits (http://www.ettelaat.com/new/). As a consequence all texts become a mix of character encodings and the occurrences of such mixed in on-line materials can affect the accuracy of natural language processing and also make the search through texts difficult.

Pre-processing task, despite its fundamental importance, is often neglected in natural language processing. Therefore we aimed to develop an open source text normalizer for Persian named PrePer to treat different cases of writing styles with various encodings.

2. PrePer: Preprocessor for Persian

PrePer (Seraji et al., 2012) is an open source software program developed in the program language Ruby for the task of editing and cleaning up texts in Persian. The program is using the existing Virastar module for some formatting tasks (Bargi, 2011). The present PrePer handles miscellaneous cases and performs functions to normalize texts into computational standard script. PrePer via Virastar takes care of the occurrences of mixed character encodings. By normalizing texts all letters in Arabic style with Arabic character encoding are edited to Persian style with mapping to Persian character encoding . Furthermore, Arabic and Western digits are all converted to Persian digits. PrePer also treats cases that Virastar is not able to treat, such as in the following cases where whitespace can deterministically and unambiguously be identified as token-internal, it is instead replaced by zero-width non-joiner (ZWNJ) in order to create a single token:

- nouns and plural suffixes ها- /-hâ/ , ان- /-ân/, ات- /-ât/, and ین- /-in/
- the suffixes /-i/ ی- or یِ- (after long vowel /u:/) when denoting indefiniteness or abstractness, as well as the indefinite suffix ای- (after silent h) and any nouns when forming indefinite nouns or abstract nouns
- nouns and pronominal clitics
- past participle verbs and copula enclitics
- nouns and verbal stems in compound words
- verbal stems and the suffix اک- /-âk/
- verbal stems and the suffixes ار- /-âr/ or گار- /-gâr/ when forming nouns of action
- nouns and their adjacent suffixes when forming adjective-adverbs or adjective-nouns
- the negative prefixes نا- /nâ-/, and بی- /bi-/ (-im, -in, -un, -less) and their adjacent word

The process has been done according to a distinct pattern using rules specified with regular expressions to search and manipulate text. Moreover, the software is flexible to be updated with additional rules for treating of further special cases.

3. Conclusion

Pre-processing a text is often performed as the first step of natural language processing tasks. Due to the use of diverse Unicode characters in Persian typing system and different writing styles, this task is a prerequisite phase of processing Persian and it is nontrivial. In this paper we introduced a solution for normalizing Persian texts through a freely available tool named PrePer.

4. References

Allen A. Bargi, 2011. *Virastar*. https://github.com/aziz/virastar.

Behdad Esfahbod, 2004. *Persian Computing with Unicode*. The Farsi Web Project (http://www.farsiweb.info).

Zina Saadi. 2007. Arabic, Farsi and Urdu Text Normalization for Natural Language Processing. *Government Users Conference*.

Mojgan Seraji, Beáta Megyesi, and Joakim Nivre. 2012. A basic language resource kit for persian. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.