DESIGN AND IMPLEMENTATION OF A SOFTWARE SYSTEM FOR DETECTING ORTHOGRAPHICAL OR MORPHOLOGICAL ERRORS IN PERSIAN WORDS

H. Hassanpour*

Department of Electrical and Computer Engineering, Noushirvani Institute of Technology
University of Mazandaran, P.O. Box 47144, Babol, Iran
h_hassanpour@yahoo.com

Mohammad Reza Razzazi

Department of Computer and Information Technology
Amirkabir University of Technology, Tehran, Iran
razzazi@ce.aku.ac.ir

*Corresponding Author

(**Received: March 14, 2007 – Accepted in Revised Form: May 31, 2007)**

Abstract This paper presents a new method for analyzing words in the Persian language context to find orthographical and structural errors regardless of the meaning. This technique tokenizes each word in a statement then tries to detect the kind of word, and analyses its correctness in terms of orthography and morphology by means of a lexicon. It should be noted that some words in the Persian language have the same stem, which are constructed by adding particles to them according to certain rules. For these words the researchers present a new method to reduce the size/volume of the lexicon and to quicken in searching.

Keywords Orthographical Error, Lexicon, Stem

چکیده این مقاله در صدد ارائه روشی برای بررسی خطاهای املایی و ساختاری کلمات در متون فارسی – بدون توجه به مفهوم جمله – می‌باشد. این سیستم با دریافت جمله ای از یک متن، تک تک کلمات آن را جدا نموده و به کمک واژگان سعی در شناسایی نوع کلمه و بررسی صحت املایی و ساختاری آن می‌نماید. در این مقاله با توجه به این نکته که بسیاری از کلمات در زبان فارسی دارای ریشه مشترکی هستند که با افزودن اجزایی به آن براساس قواعد دستور زبان بدست می‌آیند، روش جدیدی برای کمتر کردن حجم واژگان و تسریع در عمل جستجو ارائه شده است.

1. INTRODUCTION

With the advance of natural language processing techniques and expanding use of computers, investigations are performed in many languages to find orthographical and structural errors in a text [1-3]. Testing orthographical correctness and morphological consistency of words is propound as one of the applications of natural language processing [4]. Studies show that little research has so far been undertaken in computationally analyzing the Persian language [5].

Computational lexicon is among the most important resources that is needed to design a system that checks the orthography and morphology of words [6]. In a language with a rich morphology, such as Persian and Arabic, the lexicon is expected to provide enough information to enable the system to process intricately inflected forms correctly. In such a system since all information about individual words should be obtained from the lexicon, careful design of the lexicon is crucial.

The software system presented here detects

context-independent misspellings and checks the morphological consistency of words in the Persian context, and provides isolated-word error correction. The system assists the user by offering a set of candidate corrections that are close to the incorrect word. For example, when the system receives the word "پصر", it not only recognizes the misspelling, but also suggests "پسر" as a replacement. In addition, if a word like "کتابان" appears instead of "کتابها", the system detects a morphological error, and gives an appropriate message to correct it. Figure 1 represents a block diagram of the system.

It is worth noting that the results of this research may be used for designing a similar system for the Arabic language as Persian and Arabic languages have similarities in this respect. In this work, extensive research has been done on Persian grammar addressed in several Persian grammar books, such as those in [8,9].

2. PARSING AND IDENTIFYING THE WORDS

The presented system in this paper isolates words in a text using the blank space between two consecutive words. Then it evaluates the orthographical and morphological correctness of the words by means of the lexicon. If the system can find the exact word in the lexicon; it confirms the orthography and morphology of the word. Hence, the more the words in the lexicon lead to more accuracy for the performance of the system.

Following the above-mentioned idea, all derivatives of a word in the lexicon are needed. This point causes the size/volume of the lexicon to be dramatically large. Therefore, in this paper a new method for optimizing the size of the lexicon and hence improving the system's performance is presented.

3. IMPLEMENTING THE LEXICON

To reduce the size of lexicon, the set of words which can be extracted from the same stem are all replaced by just the stem within the lexicon. In order to obtain all of the derivative words from a stem existing in the lexicon; the morphological information for each of the stem words should be there. Hence, a code is inserted in front of each word containing information regarding its grammatical characteristics.

For each word it was found that an eight-bit code is sufficient to store all of its morphological information. Designing such a code system is a subtle task that is explained below.

Within Persian grammar [7,8], words can be classified into seven morphological groups: noun,

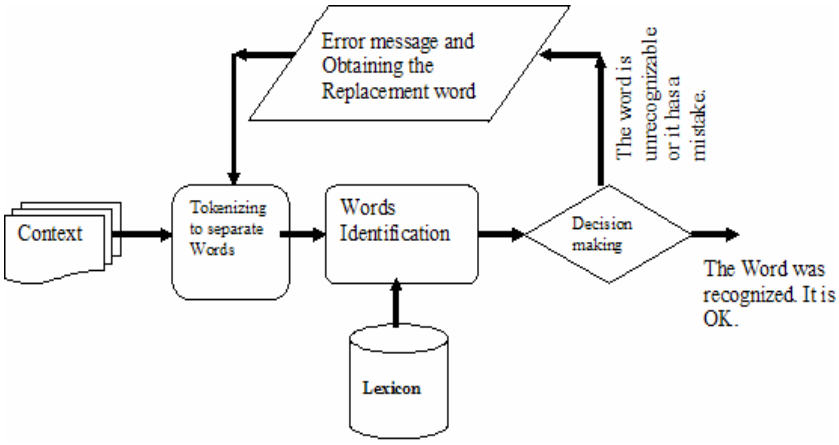

Figure 1. Diagram of the proposed system.

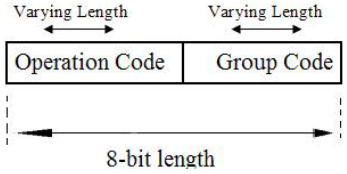

Figure 2. Format of the characteristic code for words in the lexicon.

verb, preposition, adjective, adverb, pronoun, acoustic interjection (sound). This classification has been used because different grammatical groups have their own rules to produce cognate words. For instance, in the Persian language, verbs and sounds can't be in plural form, but common nouns may appear in plural form. In addition, common nouns like "کتاب" can be pluralised into "کتابها"; "دوست" can be pluralised into "دوستان" and"دوستها"; and "امتحان" can be pluralised into "امتحانها" and "امتحانات". This kind of information must be provided by the characteristic code of the word in the lexicon.

The characteristic code is designed in such a way that it contains two parts: the first part indicates the group (type) of the word, and the second part indicates permissible operations that can be implemented on the word (see Figure 2). The characteristic codes have a fixed length, equal to eight bits; but the length of its two constituent parts varies depending on the word group. As the number of grammatical rules applicable to different word groups may not be the same, the second part of the variable code does not need a fixed length. For example interjections have the same grammatical feature in the Persian language but nouns and/or verbs have different grammatical features [7]. Thus, for recognizing interjections, only one code is enough, that is, the length of the operation code for interjections is zero.

Consider as an example, verbs can be transitive or intransitive and some intransitives can be transformed to transitive verbs on the basis of their rules [9]. In addition, there are different ways to transform a present root form to an infinitive.

Different rules must be used to make infinitive words like "دویدن", "رفتن", "شنیدن", "آزمودن" and "آموختن" from their present root verbs "دو", "رو", "شنو", "آزما" and "آموز" respectively. These

TABLE 1. The Binary Characteristic Codes for Different Types of Words. the Xs Represent the Operation Code.

Type of Word	Characteristic Code
Noun	XXXXXXX1
Verb	XXXXXX 1 0
Adjective	XXXXX 1 0 0
Pronoun	XXXX 1 0 0 0
Preposition	XXX 1 0 0 0 0
Adverb	XX 1 0 0 0 0 0
Interjection	0 1 0 0 0 0 0 0

examples show that verbs and nouns in lexicon need more than one code. Table 1 illustrates codes which are used for seven groups of words in this system. In this table, codes are in binary, and "X" indicates that the related bit can be "1" or "0" in which the former shows a particular grammatical feature for the word. For example, a code like "00100000' is used for the proper adverbs such as "never" and "sure"; and the code "01100000" is used for the common adverbs such as "year" and "time". Since common adverbs may appear as plurals, contrary to proper adverbs; this distinction has been made. It should be noted that there are different types of adverbs in Persian grammar. However, they are, from a morphological point of view, classified into two groups: proper adverbs and common adverbs. Hence, the two-bit pattern is enough for this coding.

Therefore, if the proposed system received a word in a sentence that appears as a plural and has been introduced as a proper adverb by the lexicon; from this system's point of view, the word contains a structural mistake.

4. MORPHOLOGICAL ANALYSIS AND ORTHOGRAPHICAL MISTAKES

When the proposed system finds a word in the

lexicon, regardless of its meaning in the context, it considers the word as correct in orthographical and morphological respects. Otherwise, the system supposes that the word is a kind of extended word (a kind of word that its stem exists in the lexicon), hence, it tries to detect the stem. In detecting the stem, the system may need to pass through a few steps. In each step the possible added prefixes and/ or suffixes are removed from the front and/ or back of the word respectively, until the word can be found in the lexicon. Then the characteristic code of the detected word is used to judge whether the morphology of the original word is acceptable. In other words, if the characteristic code does not let the stem have the specified prefix or suffix, the system announces that "The word has a morphological error".

For instance, suppose that, the system receives the word "کتابان". Initially it searches for the word in the lexicon, and doesn't find it. It then recognizes the "ان" from the end of the word as a sign of a plural form in some nouns in the Persian language; and hence searches for its stem "کتاب" in the lexicon. According to Persian grammar, the sign of plural is "ها" for this word. Hence, the system first prompts "This word has a problem about its plural", and then suggests the "ها" as a replacement.

Finding the stem of extended words involves a number of steps of which each step belongs to one grammatical group where the system tries to find incorporated morphemes. For example, in the step that belongs to the noun group, the system tries to find morphemes such as [ها,ات,ان, or م,و,ي,د,يم,يد,ند] or a compound form of them on the end of the word. Then it deletes those morphemes and searches for the rest of the word in the lexicon. The system will go to the next step; if the word doesn't have any morphemes related to the current group, or by deletion of the morphemes the system can't find the word in the lexicon. For example, Table 2 illustrates samples of different words when their morphemes are deleted from the beginning and/or end of the word to obtain the resulting stems.

If the system can't determine the word in any of the above steps, it will be considered from the viewpoint of existence an orthographical mistake. Since the orthographical mistakes are presented in the homophone letters; they can be classified into [ع,آ,ا], [غ,ق], [ط,ت], [ث,ص, س], [ز,ذ,ض,ظ], and [ه,ح]. If any of these letters are present in a word; they will be transformed to other homophone letters from the same group and then that word is searched in the lexicon.

5. IMPLEMENTING THE SYSTEM

The system introduced in this paper; has been implemented by the C programming language. The lexicon file used in this system contains more than 12000 words' stem. A logical record is considered for each word in the lexicon. Since the length of different words may not be the same, records with varying length are considered in the database. To speed up the searching, a three-level index has been used for the lexicon [10]. In this system a user can retrieve or add a word in the lexicon.

To evaluate the performance of the proposed system, texts with orthographical and/or morphological mistakes in their words, have been presented to the system. Except in cases that stem of words didn't exist in the lexicon, in all cases, the system could successfully recognize any morphological or orthographical mistakes.

It should be noted that, in implementing this system, it is supposed that in a word with an orthographical mistake; only one letter has been replaced with one of its homophone's letters. For instance, the word "دست" may appear as "دسط" or "دصت"; not "دشط", or "دصط".

6. CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORKS

In this paper, a system has been presented that finds the orthographical and structural mistakes in context, and presents a suggestion for retrieval. This system can be used as a preprocessor in systems which assess the structural correctness of a Persian sentence, such as the one in [5]. In fact, the proposed system can be used to make sure that contexts without orthographical and morphological mistakes in their words are fed to assess their structural correctness.

To improve the performance of this system,

TABLE 2. Samples of Different Words When Their Morphemes are Deleted from the Beginning and/or end of the Word to Obtain the Resulting Stems.

The Extended Word	Type of Word	Morpheme	Stem
کتابها	Noun	"ها"	کتاب
دوستان	Noun	"ان"	دوست
امتحاناتم	Noun	"اتم"	امتحان
دوستی	Noun	"ی"	دوست
میخورم	Verb	"می","م"	خور
مینوشیده ام	Verb	"می","ه ام"	نوش

more words can be added to the lexicon. This system can retrieve the orthographical mistake in a word if only one letter of the word has been substituted by one of its homophone's letters.

In this system, the function that extracts stem words passes consecutively through several steps, each of which belongs to one grammatical group of words. To find the root of words one can employ a searching graph furnished with estimating the cost of traversing the nodes. In this graph the chosen nodes may depend on the morpheme attached to the word.

7. REFERENCES

1. Alnajem, S., "A computational approach to the variations in Arabic verbal orthography", ***Computer Speech and Language***, Vol. 19, (2005), 275-299.
2. Oflazer, K. and Inkelas, S., "The architecture and the implementation of a finite state pronunciation lexicon for Turkish", ***Computer Speech and Language***, Vol. 20, (2006), 80-106.
3. Razzazi, M. and Hassanpour, H., "Farsi Syntax Analyser", Technical Report, ***Amirkabir University of Technology***, Tehran, Iran, (1992).
4. Sabourin C. F., "Computational Text Understanding: Natural Language Programming", Argument Analysis, Infolingua inc., Montreal, Canada, (1994).
5. Hassanpour, H. and Razzazi, M., "Design of an intelligent system for analysing the structure of Persian statements", ***Conference on Intelligent Sysytems (CIS2005)***, Tehran, Iran, (2005), (in Persian).
6. Joseph, D., and Farghaly, A. "Roots and patterns vs. stems plus grammar-lexis specifications: on what basis should a multilingual lexical database centered on Arabic be built", ***IX Workshop on Machine Translation for Semitic Languages***, New Orleans, U.S.A, (2003), 1-8.
7. Mace, J., "Persian Grammar", Routledge Curzon,

London, England, (2002).
8. Anvari, A. and Givi, H., "Persian Grammar", Fatemi Publications, Tehran, Iran,Vol. 1, (1993).
9. Anvari, A. and Givi, H., "Persian Grammar", Fatemi Publications, Tehran, Iran, Vol. 2, (1994).
10. Tsotras, V. J., Manolopoulos, Y. and Theodoridis, Y., "Advanced Database Indexing", Kluwer Academic Publishers, Boston, Massachusetts, U.S.A., (2000).