زبان تخصصی مهندسی کامپیوتر

درس چهارم:

Designing Software for Improving Writing Literacy

Part I- Writing Development: Supporting topic sentence

- در این درس، به روش دیگری از حمایت جمله نخست پاراگراف توسط جملات بعدی آن اشاره میشود. این روش بنام Enumerators میباشد. در این روش، پس از بیان کلیات در جمله اول، مثالهایی از ایده کلی در ادامه آورده میشود.

- بعنوان مثال جمله محوری ممکن است در مورد معرفی رشته کامپیوتر باشد، و در جمله دوم نرم افزار و سخت افزار به عنوان اجزاء این رشته ذکر شوند. یا جمله محوری ممکن است در مورد معرفی یک دانشگاه باشد، و به دنبال آن در جملات دیگر به معرفی دانشکده های آن بپردازد.

- Computer (General idea)
  - Hardware& Software (parts)
- University (General idea)
  - Humanity and social science college, Law and political science college, Engineering college, Art and Education college (parts)

Part II-Vocabulary

Orthographical (adj) املائی

- The conventional spelling system of a language.

Cognate (adj) هم ریشه

- Linguistics (of a word) having the same linguistic derivation as another; from the same original word or root

Transform (V) تبدیل کردن ، تغییر دادن

- Make a thorough or dramatic change in the form, appearance, or character of*: lasers have transformed cardiac surgery*

Contrary (adj) برخلاف، مغایر

Opposite in nature, direction, or meaning*: he ignored contrary advices.*

Classify (adj) **دسته بندی کردن**

- Arrange (a group of people or things) in classes or categories according to shared qualities or characteristics*: mountain peaks are classified according to their shape.*

Detect (v) **تشخیص دادن**

- Discover or identify the presence or existence of*: cancer may soon be detected in its earliest stages.*

Pop up (v) **ناگهان ظاهر شدن**

- Appear or occur suddenly and unexpectedly*: these memories can pop up from time to time.*

Substitute (v) جانشین شدن، جایگزین کردن

- Use or add in place of*: dried rosemary can be **substituted for** the fresh herb.*

Replace (someone or something) with another*: he **substituted** the drugs **with** another substance.*

Part III– Reading

Designing Software for Persian Orthographical Features

Section 1: Pre-reading Questions

- Do you have any ideas how to improve writing literacy digitally?
- Did you ever read about digitally improving the Persian orthography improvement?
- What could be new quick software to indicate Persian orthographical errors?

Section 2: Reading Passage

With the advance of natural language processing techniques and the expanding use of computers, researchers have examined many languages to find out orthographical and structural errors in a text. Examining orthographical correctness and morphological consistency of words involve the application of natural language processing techniques. Little research has so far been undertaken in computationally analyzing Persian language.

Computational lexicon is among the most important resources needed to design a system that checks the orthography and morphology of words. 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 intricate inflections correctly.

ساختار زبانی - لغت نامه - ظریف و پیچیده

The software system detects context-independent misspellings and checks the morphological consistency of words in Persian language context, and provides isolated-word error correction. The system assists a 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 shows a block diagram of the system.

مستقل از متن - اشتباه املایی - جایگزین

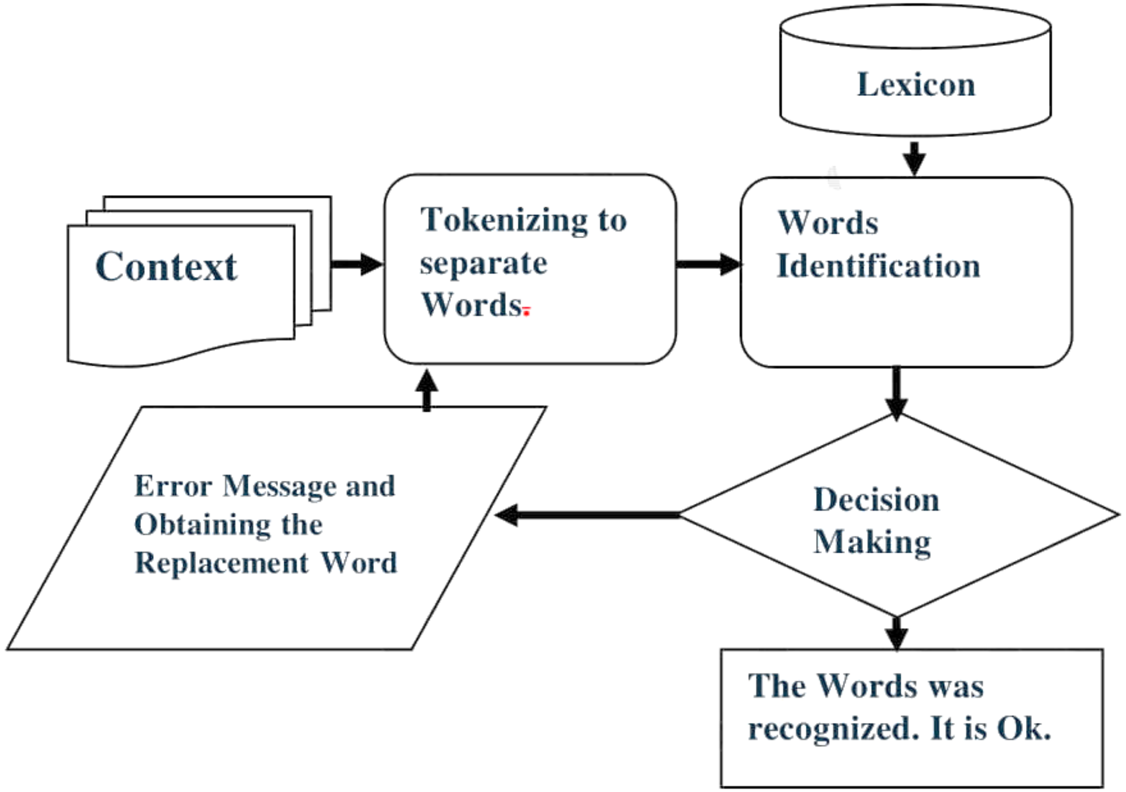

Lexicon
Context
Tokenizing to separate Words.
Words Identification
Decision Making
Error Message and Obtaining the Replacement Word
The Words was recognized. It is Ok.

Figure 1. Diagram of the system checks orthographical consistency of words in the context.

1. Parsing and identifying the words

The presented system isolates words in the 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.

Consequently, all derivatives of a word in the lexicon are needed. This point would cause the size/volume of the lexicon to be dramatically large. Therefore, a method is required for optimizing the size of the lexicon and hence improving the system performance is presented.

متوالی - بوسیله - مشتق شده - خیلی زیاد

2. Implementing the lexicon

To reduce the size of a lexicon, the stem within the lexicon replaces the whole set of words, which can be extracted from the same stem. 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.

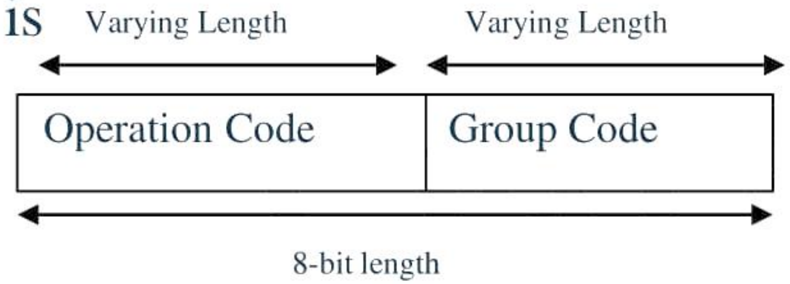
For each word an eight-bit code is sufficient to store all of its morphological information. Designing such a code system is a subtle task that is explained below.

بن کلمه - کافی - زیرکانه، موشکافانه

Words in Persian can be classified into seven morphological groups: noun, verb, preposition, adjective, adverb, pronoun, interjection. This classification has been used because different grammatical groups have their own rules to produce cognate words. For instance, verbs and interjections cannot be in plural form, but common nouns may appear in plural form. In addition, common nouns like "کتاب" can be pluralized into "کتابها"; "دوست" into "دوستان" and"دوستها"; and "امتحان" into "امتحانها" and "امتحانات". This kind of information must be provided by the characteristic code of words in the lexicon.

حرف اضافه – ضمیر - اصوات - جمع

The characteristic code 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 their 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 features in Persian language, but nouns and/or verbs have different grammatical features. Thus, for recognizing interjections, only one code is enough, that is, the length of the operation code for interjections is



Varying Length

Varying Length

Operation Code

Group Code

8-bit length

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

مجاز - تشکیل دهنده

A case in point is that verbs can be transitive or intransitive and some verbs can be transformed from intransitive to transitive ones on the basis of their rules. In addition, there are different ways to transform present root to an infinitive form.

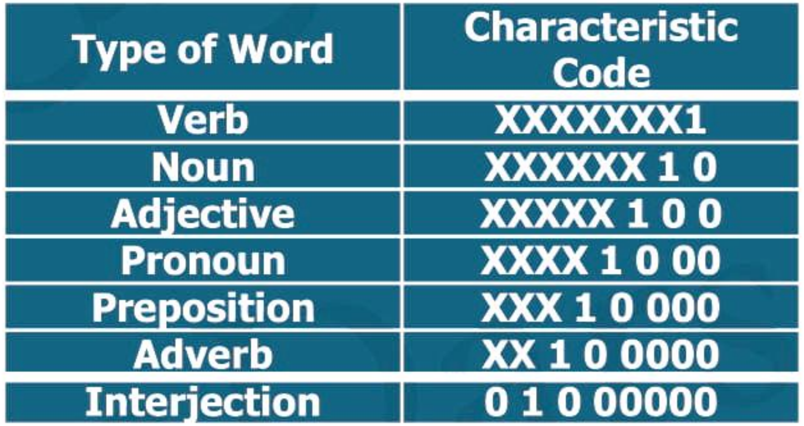
Different rules must be used to make infinitive words like "دویدن", "رفتن", "شنیدن", "آزمودن" and "آموختن" from their present root verbs "دو", "رو", "شنو", "آزما" and "آموز" respectively. These examples and the examples provided earlier indicate that verbs and nouns in lexicon need more than one code.

متعدی - بن ماضی

Table 1 illustrates codes, which are used for seven groups of words in this system. The 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 these common adverbs may appear plural, such as "years" and "times", 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 morphologically classified into two groups: proper adverbs and common adverbs. Hence, the two-bit pattern is enough for this coding. Therefore, if the system received a word in a sentence that appears plural and the word was introduced as a proper adverb, the word contained a structural error.

توضیح دادن، نشان دادن – خاص ، مناسب - تمایز

TABLE 1. The Binary Characteristic Codes for Different Types of Words. Xs represent the operation code.



Type of Word	Characteristic Code
Verb	XXXXXXX1
Noun	XXXXXX 1 0
Adjective	XXXXX 1 0 0
Pronoun	XXXX 1 0 00
Preposition	XXX 1 0 000
Adverb	XX 1 0 0000
Interjection	0 1 0 00000

3. Morphological orthographical analysis and orthographical errors

When the 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 considers the word as an extended word (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 pops up the message, "The word has a morphological error". For instance, after the system receives a word "کتابان", it initially searches for the word in the lexicon, and if it doesn’t find it, it then recognizes "ان" at the end of the word as a sign of a plural form in some nouns in Persian language. Finally, it searches for its stem "کتاب" in the lexicon. According to Persian grammar, the sign of plural is "ها" for this word. Hence, the messaging system first pops up, "This word is inaccurate in the plural form", and then suggests the "ها" as a replacement.

بودن توجه - پسوند - قضاوت کردن

Finding the stem of extended words involves a number of steps. Each step relates to one grammatical group where the system tries to find incorporated morphemes. For example, in the step that relates to the noun group, the system tries to find morphemes such as [ان,ات,ها, or ند,ید,یم,د,ی,و,م] 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 does not have any morphemes related to the current group, or by deletion of the morphemes, the system cannot find the word in the lexicon.

Finally, if the system cannot determine the word in any of the above steps, the word will be considered as orthographical errors. Since the orthographical errors are presented in the homophone letters, they can be classified into [ا,آ,ع], [ق,غ], [ت,ط], [س,ص,ث], [ظ,ض,ذ,ز], and [ح,ه]. If any of these letters are present in a word, they will be transformed to other homophone letters from the same group and following that the word is searched in the lexicon.

تشکیل دادن، جزو (چیزی) کردن یا شدن - شناسه - ترکیبی - هم صدا

4. Implementing the system

The lexicon file used in this system contains more than 12000 word stems. A logical record is constructed 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 a lexicon. In this system, a user can retrieve or add a word in the lexicon.

To evaluate the performance of the system, texts with orthographical and/or morphological errors have been tested on the system. Except for cases that stem of words didn’t exist in the lexicon, the system in all cases could successfully recognize any morphological or orthographical errors.

It should be noted that implementing this system requires a word with an orthographical error having one letter replaced with one of its homophone’s letters. For instance, the word "دست" may appear as "دسط" or "دصت"; but not "دشط", or "دصط". This system can process and detects an orthographical error in a word only if one letter of the word has been substituted by one of its homophone letters.

تسریع دادن - به جز

Part IV- Reading comprehension

Mark each statement as T (True), F (False), or NG (Not Given) to the information in the reading comprehension passage.

1. Enough research has been done in analyzing Persian language computationally.
2. The software system helps find context-dependent mis-spellings and consistency of words in Persian.
3. Code containing information on grammatical features is used for each word.
4. All transitive and intransitive verbs can be exchanged and used.
5. No difference exists between common and proper adverbs.
6. After finding the stem, the system needs to go through a few steps.
7. The first step focuses on grammatical group in cooperated with morphemes.
8. The number of word stems in this system is exactly 12000.