Presenting an Expert System for Automatic Correcting Persian Texts

Mohammad Azadnia,

ITRC (Iran Telecommunication research Center)

Summary

In this paper we have tried to present an expert system to detect and correct Persian language misspellings. It can be used to post process the texts made by OCRs or typed. Error recognition and error correction algorithms which use special heuristic functions to model incorrect words and correct them plus a Persian lexicon are the main parts of the system. To recognize errors, a lexicon which is automatically made, is used and then the most appropriate correct word is chosen. To choose them, different heuristic functions are used. Different experiments in this research have shown that use of an accurate Persian lexicon can result in great success.

Key words:

Automatic text correction, Heuristic function, Post processing, Persian language..

1. Introduction

According to the quick and increasing growth and development of information technology a huge amount of electronic texts including newspaper, web logs, internet sites, books and thesis are produced everyday. Although producing electronic documents have noticeable advantages like facilitate organizing and managing data, but the expense for generating these documents is still a deal of money. Producing these processing documents whether by typing or using an optical character recognition software, human forces and costs are required for detection the errors and correcting them. Therefore it is useful to access some expert systems to do this job automatically to decrease expenses and speeding up the production of electronic documents [1]. Some uses of the mentioned system are avoiding other errors in optical recognition system, facilitating and speeding up electronic edition in different areas like editorial electronic mail, editing input sentences of users in automatic text processing systems like machine translators, questioning and answering systems, helping users who are not good enough at the language which is being written and recommending synonyms in dictionaries [2]. The researches on application of information retrieval in documents have shown that it is not really important to pay an exact attention in documents which was created from these activities. Therefore automatic correction is very useful as long as it reaches the attention to an acceptable level. On the other hand developing automatic systems could affect decreasing expenses of producing required documents for mentioned applications efficiently. Automatically correcting written texts made from optical character recognition have been used by a lot of researchers and this area in English has improved greatly.

Modern systems like SMART and INQUERY are the ones which using statistical patterns and considering the frequencies of the words to correct them [3].

Besides researching activities various products are also found for different languages. For Persian language different efforts have made by Iranian companies such Robotic Researching Organization, Sepanta Artificial Intelligence Group and Gum Electronics which carried out the Namenegar package.

The presented expert system process in this paper is divided into three main parts: studying input document, finding probable errors and correcting detected errors. Most of the available ways to correct texts automatically have been founded by considering probable patterns and using knowledge based systems [4], [5].

The automatic correcting expert system mostly replaces the correct words with the wrong one completely automatically when it is able to choose exactly what goes out or prepares a list of recommended words for users to choose, that would be useful for the ones who are not expert enough at the written language.

Amount of similarity for these system are often introduced as a distance between two words. That is: the smaller is the distance between two words, the more probable it is to be replaced with one another. On the other hand correcting texts completely depends on how they have been produced. The texts which have been typed by human users generally have different problems from the ones created by optical character recognition software. In addition comparing the complication of Persian alphabet with English, it is much more difficult to correct Persian texts than English ones.

Automatic correcting expert systems mostly find the wrong words by a suitable lexicon consist of common words of a certain language. This lexicon includes all the roots and their derivations which language find correct [4]. However some other systems use small lexicon and a group of morphological grammar to find the incorrect words. This way to determine the correction of an input

Manuscript received March 5, 2008
Manuscript revised March 20, 2008

word, its structure is studied unless the word exists in the dictionary. In the following the architecture and algorithms used in automatic expert system to correct recommended words would be explained and the result of different texts would be given later.

2. Automatic correcting expert system

Automatic correcting expert system is used for two different texts, typed texts and the ones done by some systems like optical character recognition. Considering the difference of the errors in these two systems the heuristic function which is used to point out the distance between two words, is totally different. These functions are used during the process of choosing recommendations after finding incorrect words.

2.1 Lexicon

As mentioned before most of the automatic correcting expert systems do their job by using a lexicon consists of common words. Preparing dictionaries can be done in two different methods. The first method is to collect frequent words from huge texts, however in the other method different words derivations are made automatically by morphological grammar.

In the first method a large group of Persian texts are collected from NEWS documents and frequency of its words would be pointed out. The words with higher level of appearance would be added to the lexicon. One of the biggest problems of these methods in Persian is due to the lack of standard alphabet. Lack of similar spelling for certain words which could mean they are also some misspellings in the first text and lack of use of similar common words and that is because of extracting information out of a certain, limited source. However the mentioned method has an advantage that is the functional use of the words added to the lexicon can change according to the level of appearance of the chosen words.

In the second method of preparing a lexicon, all the roots extracted of the language and morphological grammar are used and this way it is attempted to make words automatically out of the roots grammatically. One of the most important problems is production of uncommon words in language and a large amount of Persian word derivations. On the other hand because of some obscurity about roots and morphological grammar, it is possible to add incorrect words to the lexicon for example by studying all the morphological production methods, 260 different words are derived from a verb root, a lot of the seem to be uncommon. Both methods have been used in this project. Through experiment part it is seen the results of the automatic correcting expert system using these two methods are different and based on the used lexicon there are some misspelling in the texts. It means a word that cannot be found in the lexicon, is probably misspelled which need to be corrected. Therefore some words like proper names which never existed in the dictionary would be recognized misspelled. In the following, it is described how to correct the misspelled words.

2.2 Recognizing and correcting errors

According to the studies done by [5], human errors while typing words follow certain patterns and each error is counted based on the probability of human errors. According to the mentioned reference, most of the written human errors include dropping letters, dropping spaces, replacing two alternate letters, extra space and writing a letter in a wrong way. Such errors could be considered as exchange process which its input is a correct word and the output one is wrong. This way correction method includes reverse exchanges, until it changes the incorrect input to a word that exists in the lexicon. The reverse exchanges that algorithm should do on an incorrect word includes dividing an incorrect word in to two parts (adding space), adding two incorrect alternate words (omitting space), replacing written letters with the closet ones based one the probable typing errors chart and replacing the two alternate letters. The recognizing error and correcting algorithm is illustrated in figure (1) as it is shown in the mentioned flow chart, at first algorithm tries to look up the word in the lexicon and if doesn't succeed, the word would be recognizing incorrect, and in the following it tries to correct the word. The first step to correct the errors is to add space in an input word because it is very probable to drop space while typing texts.

For the situation in which space is accidentally added at the beginning, the first part of the word is studied. If the study doesn't result in success the next word from the space is studied and if this one is also incorrect, it is probable to add a space. So by blending two alternate incorrect words, the result is looking for in a lexicon. Failure in this step causes the use of probable changing letter chart to correct the first word. These probabilities have been counted based on the physical distance between the letters on the keyboard and points out how probable a Persian letter can be replaced by another while human typing. According to this chart all the possibilities made by conversing the input word and their probabilities are considered.

Basically the words made by change in the lexicon have to be studied and then choose a word among the ones in the lexicon which sounds more probable in exchange. It is very time consuming to look up these words in the lexicon, because there are a lot of them. So it is attempted to study a smaller group of the words instead of looking all of them up. Because of that hashing by at most three letters is used, and all the three letters string types are practically considered (32768 = 32×32×32) for example one of these string is "Esterahat" (اســـتراحت) which include these substrings like: "Ehtemal"(احتمـــال), "Ehtiat" (احتيـــاط) and "Rahat" (راحت) are collocated to the related sign. Based on these data and entrance word first all the substring with one, two or three letters are extracted, then according to the word list under each mentioned signed string, the words are collected to count the similarity of the input word to others.

Figure 1 shows an example of signed words for strings with two or three letters of the word "Estekhraj"5. Because the number of made words from strings with one or two letters is large and nonesimilar, for longer words (more than 5 letters) only the words made by strings with three letters are studied. Moreover to minimize looking up similar words, another test is also considered as a creative process. Based on this process the words which their similarity in appearance seems more than 8 letters are going to be studied by exchange process. The similarity in appearance is counted by comparing the letters of two words and considers the probability of letter replacement in the entrance word in two letter areas against the compared word through this formula.

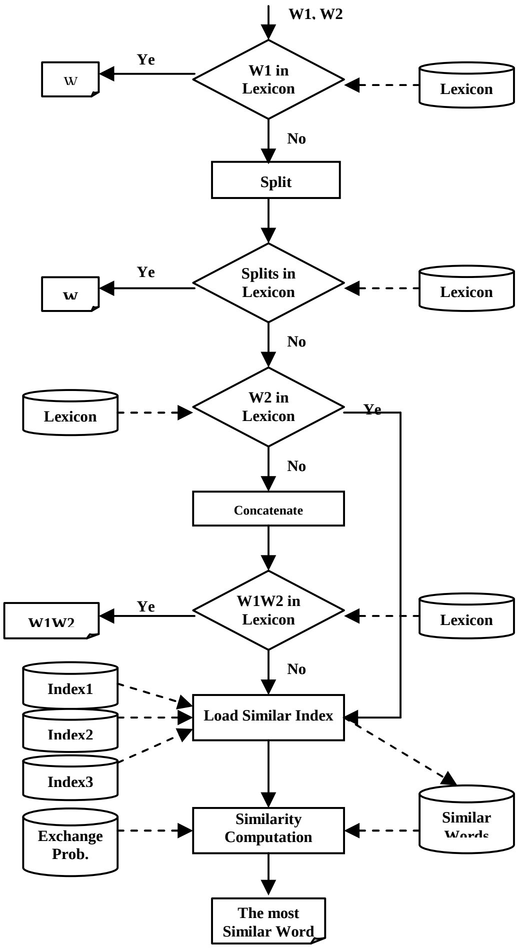

Fig. 1: The flow chart of recognizing algorithm and correcting errors.

$$letterSimilarity(wl,w2)=\frac{2*similarLeterNo(wl,w2)}{len(wl)+len(w2)} \tag{1}$$

Table 1: Example of single words with two letters

2 Similar Characters Strings	3 Similar Characters Strings
اس: اسم.تاس. ... استخراج	است: ماست.است. ... استخراج
ست: تست.شست. .... استخراج	ستح: مستحیل.استحکام . ...
تح: تحت.تحقیق. ... استخراج	تحر: متحرک.تحریک. ...
حر: حر.حرارت.محروم . ... استخراج	حرا: حراج.احرام. ...
را: رانش.مراکز. ... استخراج	راج: معراج.سراج. ... استخراج
اج: معراج.سراج. ... استخراج	

2.3 Using neighboring words

For paying more attention to choose correct words to correct errors, beside local information about a word (or at most two words next to each other) another modern pattern is also used in which texture is used, too. In this pattern it is imagined that the words dependence to the text becomes the pattern. Actually if while correcting a wrong word, its neighboring words are used, it would really help to choose the exact word. In this system for using the neighbor information and considering the effect of the next words, the n-gram pattern for Persian words has been used. This may be after finding the similar words to the one which has been recognized incorrect. The most similar one would be chosen and studied in n-gram pattern accompanied by other words. Therefore the similar word with the most probability of appearance next to other words in the text would be introduced as the correct word. In the actual system like many other patterns [6] the 3-gram has been used. However because some triplet-compounds don't exit in instructional texts like many other systems that have used the mentioned pattern [2], smoothing methods using other n-gram patterns (that is 1 and 2-gram) have been used.

In this project vertical compound of bi-gram and uni-gram with smaller coefficient accompanied by probable tri-gram pattern have been used.

$$P(W_3|W_2W_1)=\lambda_3Tri[W_3,W_2,W_1]+\lambda_2Bin[W_3,W_2]+\lambda_1Uni[W_3] \tag{2}$$

Because of avoiding correction with less similar words a limitation has been told for the amount of similarity of the words. For the words which seem incorrect to the system, no replacement is introduced.

These words are there types: the word is correct, the word is incorrect and the right user can recognize it, or the word is incorrect but the user cannot recognize the correct word which could be a proper name. For solving this problem like other communicative systems with users, the correct information is asked from user and next times based on the asked information, decision is made.

3. Experiment and evaluating results

As it was mentioned in the previous sections, beside the recognition and correction algorithm, a correct complete dictionary is one of the most required things for an automatic correcting machine. To prepare this dictionary both methods are used. For the first methods with the use of extract roots out of the Persian corpus [1, 8] and the Persian morphological rules, a collection including over 200/000 words was produced. As it was expected, although this lexicon was complete, a lot of incorrect words were found in it. So the result of the algorithm test was not satisfactory and often failed because of entering incorrect words. In the next step based on a collection of news from Islamic Republic of Iran News Agency (http://www.irna.ir), over 210.000 Persian news were picked up. For this job the words which appeared more than certain times, were collected. The results of different tests on this lexicon point out the existence of limitation among the system sensitivity and the least assigned frequency for the words of a lexicon.

As it is observed in figure (2) with the increase of the least frequency the dictionary sensitivity goes up and the number of words decreases. This method somehow improves the result. But at the end of the chart extreme reduction of the words in a lexicon follows the excess of frequency and this way the corrective machine sensitivity decrease.

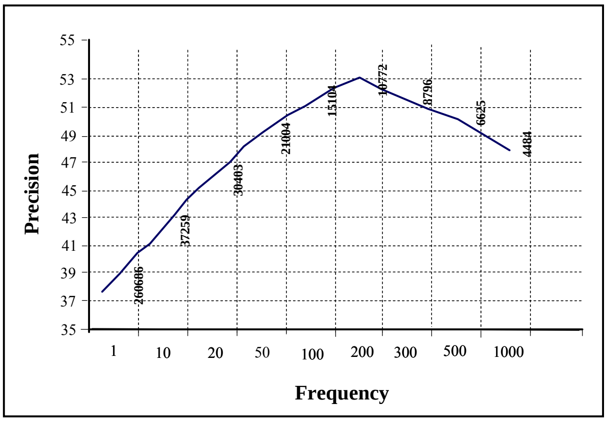

Fig. 2: The changes in system sensitivity related to the least frequency of words repetition (the numbers on the chart represent the number of lexicon words for a certain limited repetition.)

Table 2 compare the function of algorithm while using IRNA lexicon with different frequencies, the lexicon made from the corpus and the lexicon made by sharing corpus and IRNA .

Table 2: Characteristics of experiments on IRNA data and corpus

Source	*Repetition frequency*	*The number of lexicon words*	*Sensitivity percentage*
IRNA	1	260686	37.7
IRNA	10	37259	42.3
IRNA	20	30403	46.1
IRNA	50	21004	48.9
IRNA	100	15104	51.4
IRNA	200	10772	52.9
IRNA	300	8796	51.7
IRNA	500	6625	50.3
IRNA	1000	4484	47.8
IRNA - Corpus	10	31300	49.5
Corpus	1	208213	38.3

By studying the result of different tests, the mentioned system could reach the sensitivity of 52.9% in recognizing and correcting textual errors. In order to have an accurate evaluation of the system, as many natural language process usages, "Test-data" and "Gold-data" are needed. Since it is impossible to produce such great amount of text, to produce experimental data, fault injection method, which is common in other scientific areas, is used. For this reason common textual errors which have been introduced by [4], were modeled an injected in to a correct Persian text. As a result a collection of experimental data in which human errors are modeled, was prepared.

4. Conclusion

In this paper the results of researches and experiments have been presented to introduce an automatic correcting expert system. Different experiments in this project have caused originating quite a lot of heuristic functions in algorithm of recognizing errors or algorithm of choosing correct words for errors. Moreover to prepare a complete dictionary by the use of which recognizing errors is done, different recommended methods from other researches were examined and the result of each method were counted. This dictionary can be useful in other natural language usage beside its corrective system use.

References:

[1] Taghva et al, 2004. An expert system for automatically correcting OCR output, Information Science Research Institute, University of Nevada, Las Vegas.

[2] Brown et al, 1993. The Mathematics of Statistical Machine Translation, Computational Linguistics.

[3] Callan et al, 1992.The INQUERY retrieval system, In Proc. Of the 3rd International Conference on Database and Expert Systems Applications, pages 78-83.

[4] Arlandis et al, 2000. "Stochastic Error-Correcting Parsing for OCR Post-Processing", Proceedings of the 15th International Conference on Pattern Recognition, ICPR.

[5] Berghel, H. L. , 1987.A logical framework for the correction of spelling errors in electronic documents, Information Processing and Management, 23(5): 477–494, September.

[6] Angell et al, 1983. Automatic spelling correction using a trigram similarity measure, Information Processing and Management, 19_4):255-261T.

[7] Mahmood Bi Jan khan, 2002.possibilities to model Persian,No.163-162, 50-51,pages 81-93,Tehran Univ. Press,Tehran.

[8] Mahmood Bi Jan Khan, 2004.The role of textual body in writing grammar: presenting a software , No .19, pages 67-48,Iran linguistic Magazine,Tehran.

Mohammad Azadnia received the B.S. degree in Telecommunication Engineering from Iran University of Science and Technology (IUST) in 1988 and the M.S. degrees in Industrial Engineering (Industrial Management) from Sharif University of Technology in 1988. He has been working in Iran Telecom Research Center (ITRC) as a Researcher and a Project Manager since 1988. He is a member of faculty and has been a lecturer in Tehran, Amirkabir, and Zanjan Universities. He has more than 20 papers in national and international conferences in the field of Information Technology and Industrial Management.