UDC 81’

Gulila Altenbek[1,2]

[1]*Institute of Computer Science and Technology, Harbin Institute of Technology, Harbin;*
[2]*College of Information Science and Engineering, Xinjiang University, Xinjiang, P.R.China*

On the Bi-gram Model Based Auto Proofreading of Non-word Errors in Kazakh Texts

This paper discusses the Bi-gram Model based auto proofreading of non-word errors in Kazakh texts. During the process of correcting, the detections of non-word errors can be realized through syllable-based Bi-gram, that is, checking the positions of syllables in the words, and the Bi-gram co-occurrence probability; the proofreading of non-word errors then can be finished by adopting minimum edit distance algorithm and Viterbi algorithm to provide candidate words for best choice. According to the experiments results, the author proves the approach of Bi-gram Model based auto proofreading feasible.

Key words: Kazakh text; non-word errors; auto proofreading.

1. Introduction

With the rapid development of publishing industry, a large number of E-publications in Kazakh language, including E-books, E-newspapers, E-mails and office E-documents, are emerging across the region. It seems very important to ensure the correctness of those materials. A well-done error proofreading system is badly needed to detect and correct errors of the Kazakh texts. And it is one of most important tasks of natural language process.

The studies of English error proofreading system have been conducted since 1960s. People have achieved a lot in this field, such as the initial spelling checker, TYPO, invented by IBM Thomas J.Watson Research Center in 1960 and the spelling checking program developed by Ralph Gorin of Stanford University in 1971. Chinese experts have studied the Chinese auto correction system since 1990s, which also made some progress, for example, the auto proofreading system HMCTC, developed by Northeast University. Until now, the auto correction systems both in English and Chinese have reached to a higher level and been fully commercialized, especially the two successful soft ware, Word Processing of Microsoft and Heima Auto Correction in Chinese [1].

The researches of auto proofreading systems of minority languages have also made some progress, such as Stem and rule based Error Auto proofreading of texts conducted by Key Laboratory of Multilanguage Information Technology of Xinjiang University and our project about the study of Kazakh language, which has grown into a very urgent task facing us.

Based on the language features of Kazakh, this paper discusses the Bi-gram Model based auto proofreading of non-word errors in Kazakh texts. And the author just considers hand-written errors and keyboard inputting errors, ignoring other types of errors.

2. Error patterns in Kazakh text

Kazakh is an agglutinative language with word structures formed by adding derivational or inflectional affixes or suffixes to root words.

The words chosen in this paper are consisted of roots or stems and affixes. Existing Kazakh language uses Arabic alphabet to write, which thus belongs to alphabetic writing system, between words there exist separative signs. The read-write of Kazakh language are consistent, whose pronunciation is based on the unit of syllable. Thus auto proofreading at word level is the focus of our study of error auto correction of Kazakh text. Judged by the existences of the strings of characters in a dictionary, error patterns at word level are divided into the following two categories;

1) Non-word error: refers to nonexistence of a string of characters in a dictionary. For example, «دامۋ «is misspelled like»دمۋ «. (It mainly refers to misspelling)

2) Real word error: refers to the existence of a string of characters in a dictionary but with incorrect contextual meaning. For example, «ءسوزسز «is made wrong with»ءسوز «

In this paper, what we discuss are not only non-word errors, but punctuation errors and numbers errors.

2.1 Kazakh Syllable

One syllable in Kazakh language is composed by a vowel surrounded by several consonants, but even one vowel could be one syllable. The basic forms of Kazakh syllables are the followings: in this case, A represents a vowel while B a consonant. (1) A (ا), (2) AB (ات), (3) BA (جانه), (4) BAB (باس), (5) ABB (ایت), (6) BABB (کسلت). In the case of borrowed or loan words of Kazakh language, there are other forms of Kazakh syllables, which are BBA (پرولەتاریات), BBAB (تراکتور), BBABB (ترانسکریپسیا), BBBAB (ستره‌لکا) and so on.

2.2 Non-word Error Patterns in Kazakh Texts

Compared with real word errors, non-word errors account for most of the errors in Kazakh texts, mainly caused by handwriting and keyboard inputting. So this paper attaches great importance to the study of non-word errors, which can be classified into several patterns.

(1) Deletion

For example, جوسارلی (Wrong) جوسپارلی (Right)

(2) Insertion

For example, سیپاتری (Wrong) سیپاتی (Right)

(3) Character shape similarity

For examples,

وزەگە (Wrong) ورەگە □Right□

قاراپاتىم (Wrong) قاراپایىم□Right□

(4) Substitution

For example, ومىرمان (Wrong) وقىرمان□Right□

(5) The switch of «Shift» key

For example, جەتی (Wrong) جگىتی□Right□

(6) Transposition

For example, قارپا (Wrong) قاراپ□Right□

(7) Insertion of space bar

For example, قوسىلعان (Wrong) قوسىلعان□Right□

(8) Deletion or Insertion of «ء»(A symbol of soft pronunciation in Kazakh language)

For examples, بارىن (Wrong) ءبارىن□Right□

ءبىلەمىز (Wrong) بىلەمىز□Right□

(9) Abbreviation

For examples, ج ك پ CCP□ب م ۇ U.N. □

ا ق ش United states of America

3. Bi-gram Model Based Error Proofreading of Kazakh Texts

Existing approaches to the error detections in English include N-gram and looking up dictionaries.

3.1. N-gram Language Model

An ***n*-gram** is a sub-sequence of *n* items from a given sequence. *N*-grams are used in various areas of statistical natural language processing and genetic sequence analysis. The items in question can be letters, words or base pairs according to the application. One can derive a probability distribution for the next letter given a history of size *n*.

For a string of word $W = w_1, w_2, \cdots, w_n$, the occurrence of $w_i \left(1 \le i \le n\right)$ is related to the preceding words. Thus the occurrence probability of the string *W* can be calculated through the following formula:

$$p(W) = p(w_1 w_2 \cdots w_n) = p(w_1)\, p(w_2 | w_1)\, p(w_3 | w_1 w_2) \cdots p(w_n | w_1 \cdots w_{n-1})$$

$$= \prod_{i=1}^{n} p(w_i | w_1 w_2 \cdots w_{i-1}) \quad \square 1 \square$$

Given a string of words $w_1, w_2, \cdots, w_{n-1}$, the probability of w_n should be $p(w_n | w_1 \cdots w_{n-1})$. The formula can be simplified by deleting some unnecessary or minor words. Suppose that w_i is only related to the preceding n words. The formula should be changed to be the following:

$$p(W) = p(w_1 w_2 \cdots w_n) = \prod_{i=1}^{n} p(w_i | w_{i-n+1} \cdots w_{i-2} w_{i-1}) \quad (2)$$

Thus an n *n*-gram of size 1 is a «unigram» size 2 is a «bi-gram» or more is simply called an «*n*-gram».

3.2 Bi-gram Probability

When the relations between words are concerned, Bi-gram Probability refers to the dependence relationship within the words. When considering dependence relationship of a string of characters $\{x_1, x_2, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_n\}$, the dependency relationship between x_{i-1} and x_i should just be estimated as well as x_i and x_{i+1}. Through the processing of Kazakh texts, if $p(x_i | x_{i-1})$ is found within its limits; x_{i-1} and x_i are connected words. Thus during the auto detections of errors, the dependence relationships will be checked by the probabilities to judge whether those words are connected.

3.2.1 Letter Bi-gram Probability

Suppose a sequence of letter $W = l_1 l_2 \cdots l_i l_{i+1} \cdots l_m$, (*W* is a word consisted of letters), l_i and l_{i+1} are connected letters of the word. In the Kazakh corpus with the total number of all letters combination N, the co-occurrence frequency of l_i and l_{i+1} is $r(l_i, l_{i+1})$ while the respective occurrence of l_i or l_{i+1} is $r(l_i)$ or $r(l_{i+1})$. Probability should be

$$p(l_i) = r(l_i)/N \text{ or } p(l_{i+1}) = r(l_{i+1})/N; \quad (3)$$

while co-occurrence probability should be $p(l_i, l_{i+1}) = r(l_i, l_{i+1})/N$. (4)

And the mutual information between l_i and l_{i+1} should be

$$I(l_i, l_{i+1}) = \log_2^{\frac{p(l_i, l_{i+1})}{p(l_i) \cdot p(l_{i+1})}}. \quad (5)$$

3.3 Minimum Edit Distance Algorithm

Existing approaches to error proofreading in English include minimum edit distance algorithm, N-gram, rule-based algorithm, statistics based algorithm, looking up dictionary and feature model based algorithm [2].

The minimum edit distance between two strings is defined as the minimum number of point mutations required to change one string to the other. For example, the distance of minimum edit distance between جوسارلى and **جوسپارلى** is 1. Three methods can be used to achieve this goal, including trace, alignment and operation list. For a Kazakh word, whose length is supposed to be n, the possible edit operations may include: n types of deletions, n-1 types of transpositions, 33n (in fact 29n) types of substitutions and 33(n+1) types of insertions.

3.4 Syllable Bi-gram Model in Kazakh Texts

Bi-gram based error auto proofreading of Kazakh texts adopt Viterbi algorithm, which a dynamic is programming algorithm (DPA) for finding the most likely sequence of hidden states. The basic idea of DPA is decomposing the problem into many sub-issues, firstly solving the most basic problem, then gradually out push to find the optimal solution of bigger sub-issue, after the limited steps to get the optimal solution of the whole problem [3].

3.4.1 syllable BI-GRAM model based on position information

The figure below is a simplified model of syllable bi-gram based auto correction of Kazakh texts; each syllable is kept at one node with a weight value on the connecting line between the nodes, which expresses the correlation rate between the nodes.

Figure 1: The frequency between syllables in a word (Note: S1 represents the first syllable, S2 the second and so forth.)

The information included at each note is quite few, which only shows the correlation rate between those two syllables, S1and S2, in the same Kazakh word. Through further analysis of the positions of two syllables in the Kazakh word, the possible different positions can be observed to get the modified figure 2. Syllable information as well as its position in the word is kept at nodes.

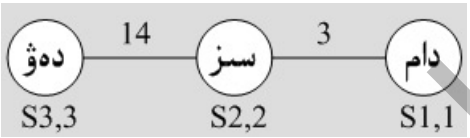

Figure 2: Syllabe positions in a word decided frequency (Note: S1,1 represents دام at the first place, S2,2 سىز at the second place, and so forth.)

In order to meet the need of searching the syllables during the process of checking and correcting, the syllable bi-gram model includes the following information: syllables, the lengths of syllables, the positions of syllables etc.

4. Syllable Bi-gram Model Based Error Detections and Proofreading of Kazakh Texts

After analyzing error checking procedures, we know that syllable Bi-gram based auto proofreading of non-word errors of Kazakh texts are divided into two categories:

1). the word can be segmented correctly but with low syllable bi-gram probability.

2). the word can not be segmented correctly, which is mainly caused by lack of vowels [4].

4.1 The Statistics of Syllables in Kazakh Language

Suppose a string of words $W = w_1 w_2 \cdots w_{n-1} w_n$, the syllables of the string amount to $W = s_1 s_2 \cdots s_{m-1} s_m$ $m < n$. The first syllable is s_1, the second is s_2, and so on. There are 6 syllable forms in Kazakh language. Based on the six syllable forms, 60,230 words in Kazakh Stem Dictionary can generate 128,271 syllables within 3600 syllable categories.

With reference to the positions of syllables, the list below gives the categories and the number of syllables.

Table 1

The statistics of syllables

The positions of syllables	The number of syllables	10 various syllables
1	2375	722
2	2194	718
3	1276	394
4	605	150
5	237	48
6	92	17
7	34	1
8	7	0
9	3	0

Based on the statistics of Kazakh corpus, much used syllables database and their collocation collections have been established.

4.2 The Real Word Dictionary of Kazakh Language

This paper uses *Detailed Annotations of Kazakh Words* published by Xinjiang People' Publishing House as our research object, which covers more than 60,000 words. The Stem list consisted of 62,324, stems was established in the form of txt. The words of the dictionary are alphabetically arranged. Checking can be done by using the list. As we know Kazakh word structures formed by adding derivational or inflectional affixes or suffixes to root words, Thus every new Kazakh word can be formed by adding affixes to the old one. It is impossible for the real word dictionary to collect all the Kazakh words. The stem extraction and morphological analysis are also required to conduct for checking unregistered words in the dictionary [5].

4.3 Error Proofreading of the First Category

For the syllables with low syllable bi-gram probability, all the possible syllable forms whose distance is 1 can be obtained through minimum distance algorithm. And then the optimal solution can be gotten through Viterbi algorithm. Viterbi algorithm is based on the unit of sentence. A marking string is chosen for each sentence to finish marking. For example,

The string of characters «جوسارلی» can be correctly segmented into three syllables, which are جو1 / سار2 / لی3

Based on the minimum edit distance algorithm, candidate syllables Table, whose distances among syllables are all «1», can be stated below.

Table 2

Candidate syllables

Candidate syllables of لی	Candidate syllables of سار	Candidate syllables of جو
تی	سا	جا
لتق	سر	و
چی	пар	جوس
⋮	⋮	⋮
the total number: 43	the total number: 51	the total number:39

According to the syllable bi-gram, the candidate syllables Table is stated below.

Table 3

Candidate syllables based on syllable bi-gram model

The positions of the second syllable and the third syllable	The positions of the first syllable and the second syllable
سار لی	جو سا
سار لا	جو عار
سار لتق	جا سار
⋮	⋮
137 types of combination Frequency □0	75 types of combination Frequency □0

Based on the Bi-gram the probability of candidate syllables can be obtained and the Viterbi algorithm can be used to find the most likely sequence. The formula can be drawn as follows:

$$\delta_t(i) = \max_{q_1,q_2,\cdots,q_{t-1}} p\left(q_1 q_2 \cdots q_t = i, O_1 O_2 \cdots O_t \middle| \lambda\right). \quad (6)$$

The probabilities of candidate syllables are Table as follows:

Table 4

The Candidate Syllables based on the Viterbi Algorithm

The combination of syllables	probability
جوس پار لا	0.056140350877193
جوس پار لی	0.0210526315789474
فو نار لا	0.00142973856209151
جۆ سان دی	0.00084468377151304
⋮	⋮

Finally, the candidate words will be matched with the real words in the dictionary and the most likely candidate words Table can be obtained.

4.4 Error Proofreading of the Second Category

For the second category, after a string of characters or letters is segmented by segmenting program, the letters which are against rules will be made unnecessary. So we make it certain that the wrong syllables should be in the positions of those unnecessary letters.

For example, جگتستكتەری, the result after syllables' segmentation should be two letters:

تىس1/ تىك2/ تە3/ ری4, from the result, جگ is not in the word any more, from which we could judge the positions of wrong syllables.

Candidate syllables Table, in which the distance of every string of characters is 1, is as follows:

Table 5

Candidate syllables without syllable segmentation

Candidate syllables
جا
جی
جە
جۆ
⋮
The total number:11

Candidate syllables Table could be obtained through positions of syllables and their combinations with the following syllables.

Table 6

Candidate syllables list without Bi-gram syllables segmentation

The candidate syllable is the first syllable while تىس is the second
جا تىس
جە تىس
جۆ تىس
the total number:3

The probability of candidate syllables can be ensured by syllable bi-gram, and then the candidate words can be obtained by Viterbi algorithm based on the probability.

Table 7

Candidate syllables list based on the Vieterbi Algorithm

The combinations of syllables	probability
جە تىس	0.00799086757990868
جا تىس	0.000369822485207101
جۆ تىس	0.0015552099533437

Finally we just match the candidate words with the real words in the dictionary, on that condition the final candidate words list can be easily gotten.

5. Experimental results

5.1 Preparations for experiments

1) Preparing Kazakh corpus

The three books are chosen as experimental materials, which are 1. Kazakh Customs and Folklore (words: 23,747; Morphology: 7,774) 2. Fairy Tales All Over the World (words: 28,092; morphology: 7,791□3. The Road to Abuy (Half of the first two volumes□words: 66,111; morphology: 18,637) The total words: 117,950; the total morphologies: 28,076.

60 % of each book will be used as training texts while 40 % will be used as testing texts. About newspapers as experimental materials, 30 articles would be used as kinds of training material while other 10 articles as testing materials.

2) Evaluation index

The evaluation index of auto proofreading system

Return rate = the correctly warned mistakes
the total mistakes in the text ×100□ (7)

Accuracy rate = the correctly warned mistakes
the total number of warned mistakes ×100□ (8)

5.2 Experiment Results

The experiment results are stated as Table 8:

Table 8

The return rate and the accuracy rate of the system

The corpus	return rate	accuracy rate
Texts	80.3□	82.5□
Newspapers	77.6□	70.7□

The results show that this model is effective and feasible targeting at corpus with a large scale.

6. Conclusions and Future Study

This paper analyzes the non-word errors of Kazakh corpus, and then adopts bi-gram and looking up dictionary to detect non-word errors based on the real word dictionary, stem extraction and syllable bi-gram model. Secondly the author uses minimum edit distance algorithm and syllable bi-gram model and viterbi algorithm to provide candidate words for correcting non-word errors. And finally the paper has drawn its conclusion that the syllable bi-gram model based auto proofreading of non-word errors are practical because of its higher return rate and accuracy rate. And the experiments show that the system can ensure the correctness of the texts and reduce the labor work of working staff.

What the project should continue to do is to improve the bi-gram co-occurrence probability together with co-occurrence probability list, grammar rules database of Kazakh language, and the algorithm of statistics combined with language features. So we can improve the efficiency and performance of the system and further the study of real word errors and adjust and enlarge the corpus in order to make the training texts more expanded and more reasonable.

Acknowledgment

This work is being done as a part of the Project *«Technology Exploration of Constructing the Corpus of Contemporary Kazakh Words»* that is being supported by the Natural Science Foundation of China under Grant No.60763005 & No:61063025, and Project funded by Ministry of Education under Grant No: MZ115–92.

References

1 *Daniel. Jurafsky, James H.Martin.* Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. — M., 2005. — P. 116–117.

2 *Kukich K.* Techniques for Automatically Correcting Words in Text [J]. ACM Computing Surveys. — 1992, 24 (4). — P. 377–439.

3 *Lei Zhang, Ming Zhou, Changning Huang*, etc. Automatic Detection and Correction of Typed Errors in Chinese Text [J]. Applied Linguistics. — 2001. (1) — P. 19–26.

4 *Damerau F.J.* A technique for computer detection and correction of spelling errors [J]. Communications of the ACM, 7(3). — P. 171–176.

5 *Milat etc*. Contemporary Kazakh language. — Xinjiang People's Publishing House, 2003.

Гулила Алтенбек.

Қазақ мәтіндеріндегі тілдік емес қателердің нәтижелерін тіркейтін Bi-gram буындық бағдарламасы

Мақала тілдік емес (пунктуациялық, емле) қателердің нәтижелерін тіркейтін Bi-gram буындық бағдарлама құрастырған эксперименталдық тексеру жайындағы мәселені қарастыруға арналған. Бұл түзету бағдарламасы негізінде қазақ силлабемасы (буыны) сипаты, сөздегі силлабема позициясы және математикалық және статистикалық мәліметтерге негізделген силлабема шекарасындағы мүмкін болатын нұсқаларды шығару жөніндегі түсінік бар. Эксперименталдық зерттеу нәтижелеріне сүйене отырып, авторлар аталмыш әрекеттің тиімділігі мен жүзеге асу мүмкінігінің жоғары екендігін танытады.

Статья посвящается рассмотрению вопроса об экспериментальной проверке созданной слоговой программы Bi-gram, фиксирующей результаты неязыковых (пунктуационных, орфографических) ошибок. В основе такой программы корректировки лежит представление о характере казахской силлабемы (слога), позиции силлабемы в слове и вычислении возможных вариантов в ее границах, основанных на математических и статистических данных. Результат экспериментального исследования корпуса таких ошибок дал возможность автору убедительно показать результативность и осуществимость данного подхода.