Acta of Turin Polytechnic University in Tashkent, 2023, 30, 38-42
Published Online September 2023 in Acta TTPU (http://www.acta.polito.uz/)

SOLVING THE PROBLEMS OF NORMALIZATION OF NON-STANDARD WORDS IN THE TEXT OF THE UZBEK LANGUAGE

[1]**Ibragimova S.N.**, [2]**Turayev B.Sh.**, [2]**Abdullayeva M.I.**
[1]Research Institute for the Development of Digital Technologies and Artificial Intelligence
[2]Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Email: snibragimova@mail.ru

Abstract– Text normalization is an important component of the text-to-speech (TTS) system, and the difficulty of text normalization lies in distinguishing between non-standard words (non-standard words). In this paper, a taxonomy of non-standard words based on Uzbek speech has been developed, and a two-stage strategy for determining non-standard words has been proposed. The proposed two-stage strategy for identifying non-standard words provides an accuracy of 98.53% in the open test. Experiments show that non-standard taxonomy of words provides high initial performance.

Key words– non-standard words, taxonomy, text normalization, state machine, classification, maximum entropy classifiers.

I INTRODUCTION

In life, when working with real text for machine translation, automatic speech recognition or speech synthesis and analysis, the text always contains numbers, abbreviations, dates, currencies, etc. The text may consist of words whose pronunciation is usually not found in dictionaries or lexicons, such as “BMT”, “UzKhDP”, “TATU”, etc. Such words are called non-standard words. In principle, any system that works with unrestricted text should be able to work with non-standard words. In this case, each text document goes through a series of processing steps to standardize it. The text of the Uzbek language, in addition to ordinary words and names, contains non-standard words, including numbers, abbreviations, dates and amounts of money. As a rule, non-standard words cannot be found in the dictionary, and it is also impossible to interpret their pronunciation using the standard rules for converting the “letter-sound” transition [1-2, 13].

Non-standard words have several categories:

- numbers whose pronunciation changes depending on whether they refer to currency, time, telephone numbers, postal codes;
- abbreviations, abbreviations, acronyms;
- punctuation;
- dates, times, units and URLs.

Many non-standard words are also homographs, i.e. words with the same written form but different pronunciation:

- IV, which can sound differently: four (to‘rt), fourth (to‘rtinchi);
- three- or four-digit numbers, which can be dates and regular numbers (e.g. 2040-yil, 2040 tonna).

Non-standard words need to be normalized to their corresponding standard words, a process called text normalization. In English, numeric expressions and abbreviations are non-standard words. Even sentence segmentation is part of text normalization. For the Uzbek language, numbers, symbols and alphabets that are not Uzbek words must be normalized to the forms of the Uzbek language. Non-standard words may be replaced by other standard words depending on the local context and the genre of the text. Hence, the problem is reduced to finding complex homographs [3]. In Nuance Vocalizer, more than 20% of the main application code (code metric line) is devoted to text normalization, and new input formats continue to be added [4]. Conventional text normalization methods are based on simple rules. However such simple custom rules are difficult to write, maintain, and adapt to new domains. On the other hand, when detecting homographs, many machine learning methods are used that have shown their advantages. Decision trees and decision lists are used to normalize text in English and Hindi, as well as for Uzbek [5]. Text data is classified and used according to the support vector machine (SVM) classification algorithm [6].

However, most Uzbek text normalization modules are rule-based and run before the word segmentation process. Because in the Uzbek text spaces between words are used in different cases. In literature, he adopts the method of normalizing the Uzbek text based on an external rule. It uses over 15 external rules and verbal and speech data. Still, others put word segmentation, named object recognition, and custom word processing into a single framework.

This article proposes a two-stage strategy for identifying non-standard words in the Uzbek text. The proposed text normalization algorithm does not require a word segmentation process. This algorithm includes finite automata that identify non-standard words from the text and perform an initial classification, then classifiers with maximum entropy are used for further classification.

II THE METHODOLOGY

1. CLASSIFICATION OF NON-STANDARD WORDS

A non-standard taxonomy of words was developed following a systematic review of the extensive TTS corpus. Based on this taxonomy, a three-level normalization process was developed. Finite automata are used for non-standard word detection and initial classification. Maximum entropy classifiers are used to further classify non-standard words, and numeric state converters are used to generate standard words [7-9]. Non-standard taxonomy of words underlies text normalization. It defines categories of non-standard words, according to which non-standard words are identified, classified and modified. Arabic, Roman numerals and some symbols are the main normalized objects in the text in Uzbek [10-11].

Table 1 provides a brief description of the taxonomy of non-standard words. Non-standard words are first classified according to their format. 95% of the 276 non-standard words in the algorithm are numeric strings and various combinations of characters (period, hyphen, slash, colon, etc.). Symbols is another category to change and some symbols have multiple pronunciations. The normalization of URLs and email addresses is obvious. Strings of the English alphabet have corresponding Uzbek translations. All other unique non-standard words are also added to the "Other" category. In total, the taxonomy includes 48 types of non-standard words in different formats. Some of these species have excellent pronunciation, while others do not.

Numbers	numbers	1,2,3, ... etc....
	with a dot	1.29, 2000.9.10, 162.105.81.14, . . .
	with a hyphen	1998-2002, 2000-9-10, 4-3-2-1, . . .
	with a slash	1/3, 2000/9/10, . . .
	indicators	10:15, 10:15:20, ...
	additions	%, (ten thousand), adjectives, ...
	range	100-200 Ď (from 100 to 200), . . .
	other	'99, ...
Symbols	-, /, :, ., ×, >, =,	
Other	URL, Email, Alphabets, . . .	

TABLE 1: TAXONOMY OF NON-STANDARD WORDS BASED ON INPUT FORMATS

Non-standard words whose pronunciation is determined by formats are called basic non-standard words (BNSW), and ambiguous non-standard words are called ambiguous non-standard words (ANSW). Tables 2 and 3 below show some examples of BNSW and ANSW, respectively. Table 2 shows the proportional distribution of non-standard words in the Uzbek text. From the table, you can see that the probability of occurrence of BNSW among all non-standard words is 55%, and their number is 84% of all possible non-standard words. It follows that 84% of non-standard words are written according to the established format (for example, 30%, 10 kg, 6 yil) and only 16% are ambiguous (for example, b2b, 115, 1998-2000).

Class of non-standard words	Example	Percentage
Indicative numbers	35 P inchi,nchi	55%
Integer	100 $	8%
Percent	10%, 12.5%	6%
Date	27 oktabr	4%
Numbers and words	15 ming	3%
Number basis	5 kg, 10 sm	2%
Year	5 yil	2%
Other	Win32	4%

TABLE 2: BNSW EXAMPLES

Table 3 shows some categories of responses and possible ways to record them. It is clear that some non-standard words have a high level of ambiguity and their meaning requires internal and contextual information.

2 TEXT NORMALIZATION METHOD

To normalize the Uzbek text, an algorithm has been developed that consists of three main stages.

1. Highlighting non-standard words and preliminary

Class of non-standard words	Words	Example
Numbers	Decimal numbers	2 ga 11 (2.11 metr)
	Integer numbers	110
	Vote	110
	English alphabet	p2p
a-giper	year-year	1998-1999
	Phone number	+99893 385 34 34
	number-number	737-200 (Boying 737-200)
	Subtraction	100-1=99
Slesh	Fraction	1/3
	Date	2001/01
Dominant	Time	10:15 (10:15 soat)
	Steps	10:15

TABLE 3: ANSW EXAMPLES

classification. In the first stage, a machine learning algorithm is used to extract non-standard words from the real text and carry out a preliminary classification. At this point, the BNSW classification is completed.

2. Definition of subclasses to display the answer. To derive the answer, the result of the initial classification is used to determine the subclass. To perform this step, maximum entropy classifiers are used.

3. Generation of the Standard word. If a non-standard word is tagged with a class tag, the restricted state switch converts it to a standard word. The text normalization scheme is shown in Figure 1.

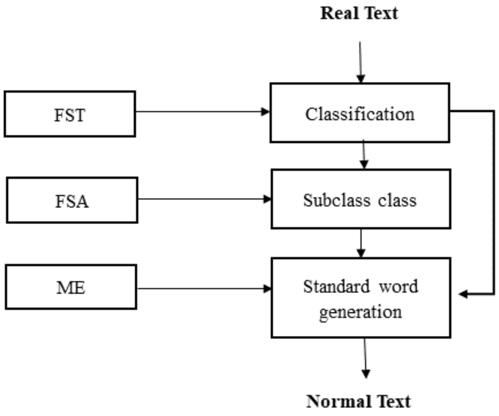


Fig. 1: Text normalization scheme

The full cycle of normalization of non-standard words is shown in Figure 2.

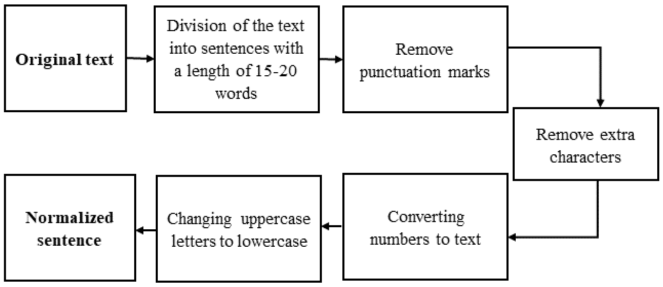

Original text
Division of the text into sentences with a length of 15-20 words
Remove punctuation marks
Remove extra characters
Converting numbers to text
Changing uppercase letters to lowercase
Normalized sentence

Fig. 2: Text normalization process.

III RESULTS AND DISCUSSION

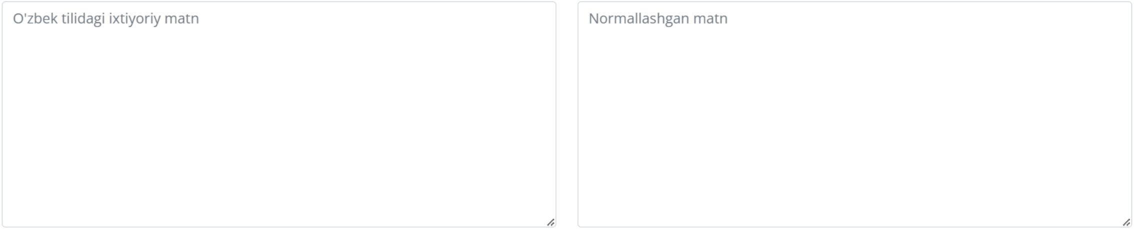
An experimental study of the performance of the proposed algorithm was carried out on the example of solving a practical problem. The system interface consists of two fields, as shown in Table 4. In the left field of the system, the source text in Uzbek is entered, and the normalized text in Uzbek is displayed in the right field. The text contained words from the BNSW and ANSW classes.

Denormalized text	Normalized text
Bugun sana 27-mart 2023-yil. Meni bugu 1-param bor.	bugun sana yigirma yettiinchi mart ikki ming yigirma uchinchi yil. meni bugu birinchi param bor.
1 kg da 1000 gr bor	bir kilogramm da bir ming gramm bor
1998-1999 yillari men maktabga borgan edim	bir ming to'qqiz yuz to'qson sakkizinchi bir ming to'qqiz yuz to'qson to'qqiz yillari men maktabga borgan edim

Figure 3-4 shows the system interface and the results of Uzbek text normalization.

IV CONCLUSION

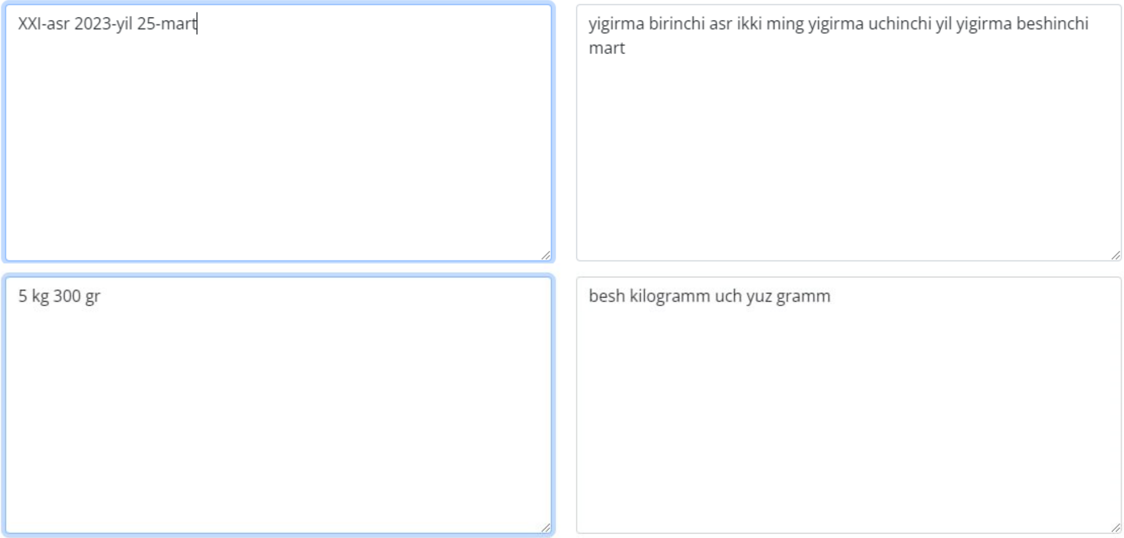
This article provides a comprehensive study of the normalization of the Uzbek text. On the basis of a large corpus, a non-standard taxonomy of Uzbek words was developed. A two-stage non-standard strategy for word classification is proposed, which is carried out using an automaton with a



O'zbek tilidagi ixtiyoriy matn

Normallashgan matn

Fig. 3: System interface.



XXI-asr 2023-yil 25-mart

yigirma birinchi asr ikki ming yigirma uchinchi yil yigirma beshinchi mart

5 kg 300 gr

besh kilogramm uch yuz gramm

Fig. 4: Result of text normalization.

finite number of states for the initial classification and classifiers with maximum entropy. Experimental results show that this approach provides good performance and generalizes well to new areas. In addition, this algorithm is based on working with symbols and does not require a word segmentation process.

V References

[1] Jurafsky D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (University of Colorado, Boulder) Upper Saddle River, NJ: Prentice Hall (Prentice), D. Jurafsky, J.H. Martin, Computational Linguistics, 2000, T. 26, N 4.

[2] Richard Sproat, Alan Black, Stanley Chen, Shankar Kumar, Marsi Ostendorf, and Christopher Richards, «Normalization of Non-Standard Words», Computer Speech and Language, 2001, 15(3): pp. 287-333.

[3] Allen, Jonathan, M. Sharon Hunnicutt, and Dennis Klatt, «From Text to Speech: the MITalk System», Cambridge University Press, Cambridge, 2001.

[4] Abdurakhmonova N. Z. “Modeling Analytic Forms of Verb in Uzbek as Stage of Morphological Analysis in Machine Translation” Journal of Social Sciences and Humanities Research. 2017, 5(03):89-100.

[5] Abduraxmonova, N. Z. “Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) il dis. aftoref.”, 2018.

[6] Musaev M. M., “Modern methods of digital processing of speech signals” Bulletin of TUIT, 2017, Vol. 2, N 42, pp. 2-13 [In Russian].

[7] Musaev M.M., Xujayarov I.Sh., Ochilov M.M., “Recognition of phonemes of the Uzbek language based on machine learning algorithms” Informatics and energy problems, 2019, Vol. 6, [In Uzbek].

[8] Alimuradov A.K., Churakov P.P., “Review and classification of methods for processing speech signals in

speech recognition systems” Measurement. Monitoring. Control. Control, 2015, Vol. 2, N 12, pp. 27-35 [In Russian].

[9] Musaev M., Khujayorov I. and Ochilov M., “The Use of Neural Networks to Improve the Recognition Accuracy of Explosive and Unvoiced Phonemes in Uzbek Language”, Information Communication Technologies Conference (ICTC), Nanjing, China, 2020.

[10] Musaev M. M., Rakhimov M. F. “Algorithms for parallel processing of speech signals” Bulletin of TUIT, 2018, Vol. 2, N 46, pp. 2-13 [In Russian].

[11] M.M. Musaev, U.A. Berdanov, K.E. Shukurov, «Hardware and software solution signal compression algorithms based on the Chebyshev polynomial» International Journal of Information and Electronics Engineering, 2014, t. Vol. 4, N 5, pp. 380-383.

[12] Jalil, Masita, Ismailov, Alisher and others The Development of the Uzbek Stemming Algorithm. Advanced Science Letters. 2017, pp. 4171-4174.

[13] Sproat, Richard, editor, “A Computational Theory of Writing System”, Cambridge University Press, Stanford, 2000