ISSN 1512-1232

NATURAL LANGUAGE PROCESSING IN THE STATE DECENTRALIZED SYSTEMS OF GEORGIA

Irakli Kardava[1,2], Elene Esiava[1,3]

[1]Sokhumi State University, 61 Politkovskaya street, 0186,Tbilisi, Georgia
[2]Adam Mickiewicz University in Poznań, ul. Wieniawskiego 1, 61-712 Poznań, Poland
[3]Maastricht University, Minderbroedersberg 4-6 6211 LK Maastricht, The Netherlands

irakar@amu.edu.pl
elene.esiava@sou.edu.ge

Abstract

As we know, the State Public Registry uses Blockchain Technology in Georgia. Naturally, the document turnover proceeds in the Georgian language. However, due to the peculiarities of this language and, also, the small number of Georgian-speaking people, there has not been made significant progress in the computer processing of the Georgian language, so far. The present paper investigates the implementation of well-known machine learning algorithms for the Georgian language in the public registry using the blockchain system (this, in turn, will facilitate the complete computer processing of the Georgian language (NLP) and use it for other purposes as well). However, it should be noted that the Georgian language differs from other languages a lot, which makes it necessary to modify the algorithms in order to successfully work, for example, for English or other already processed languages, to the needs of Georgian with its grammatical peculiarities. Moreover, it should be noted that the use of the approaches proposed by us in this paper further accelerates the work of this algorithm at the expense of reducing iterations and saving computing resources. A more general purpose of the research is to make it possible to classify texts in the blockchain system the existing documents in registry repositories, identify minimal editing distance between words, calculate the probability of probable sequences between given words, and so on.

Keywords*: Blockchain Technologies, Machine Learning, State Public Registry, NLP*

1. Introduction

Georgia is the first country in the world that successfully implemented blockchain technology in real estate registration. The National Public Registry Agency was the first to use blockchain technology in land registration in February 2017. The use of blockchain technology makes registration operations accessible, transparent, and even more secure around the world, making this service free and convenient for citizens.

Merging blockchain with AI-based NLP supports confidentiality gaining huge attention from the global audience. The main features of blockchain such as fault-tolerant, immutability, and irreversible made it an ideal to carry the sensitive data using NLP algorithms. The decentralized architecture gives us a reliable database for the data. So, Blockchain is an ideal technology for AI-based NLP systems. Blockchain and NLP together support confidentiality in organizations when it comes to securing sensitive data. ML algorithms are used on the Ethereum blockchain allowing machines to learn from data stored in a blockchain (decentralized database). This gives them the possibility to manage their own data. The blockchain has many applications that have been used in machine learning and natural language processing, such as a smart contract that could be made in

ISSN 1512-1232

order to automate payment processes between two individuals also between two companies and it ensures the confidentiality of the transactions[1-3].

In order to classify hundreds of thousands of documents located in a decentralized system considering various content requirements, it is not enough to simply use standard mechanisms. it is necessary to use more complex - intelligent algorithms, on the basis of which it will be possible to determine the semantic meanings of verbal texts by giving different contexts, to calculate the percentage similarity between the given words. To detect the probability value of their possible sequences also, to construct additional information from the given text, besides the traditional one, for carrying out further intellectual manipulations, get precise answers with higher accuracy to the needed request create daily evolving intuition models, that will let make predictions with the highest accuracy. One of the main instruments for achieving this goal is the word: morphological composition and semantic meaning of the connection of the alphabetical characters.

on the one hand, the novelty of our approach suggests the Georgian State Registry system should obtain machine learning components and mechanisms, which, in turn, will be processed by the Blockchain system.

Here, it is implied that already existing text documents will be used as the needful information necessary for educating and training; and the documents created in the future will serve as a means of developing (modifying) potential models. It is common knowledge, that given language X undergoes changes over time, even at the least. The old norm might be replaced by a new one, or both may coexist and etc. That is, the model will be "self-developing", which is flexible to this type of innovation. On the other hand the novelty, is that the principle of operation of such algorithms (in the case of computer processing of the Georgian language [5-7]) within a decentralized system will be differ technically from their classical nature [4].

2. Existing NLP algorithms and their modification taking into account the specifications of the Georgian language

2.1 Minimum Edit Distance

To find the most correct candidate wcorrect with the wrongly given word werror, first of all, it is necessary to calculate the minimal editing distance of the word. This approach requires checking all the members of the word block (a unique list of words). That is, if V=1000000, then the major cycle's iteration is i=V=1000000, too. Finding the minimal editing distance could be done by following a classical algorithm (1) [1]:

$$D[i,j] = min\begin{cases} D[i-1,j] + del - cost(source[i]) \\ D[i,j-1] + ins - cost(target[j]) \\ D[j-1,j-1] + sub - cost(source[i], target[j]) \end{cases} \quad (1)$$

In this case the value of every single activity is: insert=1, delete=1 and substitution=1; also, the number of minimum editing steps = minimum sum of the weights of the editing activities. This is because the weight of each of the three activities equals to one another. There is an approach where the replacement activity = 2. In this case accordingly, the third branch of the algorithm looks differently. See Formula (2) [1]:

$$D[i,j] = min\begin{cases} D[i-1,j] + del - cost(source[i]) \\ D[i,j-1] + ins - cost(target[j]) \\ D[j-1,j-1] + \begin{cases} 2; if\ source[i] \neq target[j] \\ 0; if\ source[i] = target[j] \end{cases} \end{cases} \quad (2)$$

ISSN 1512-1232

The principle of its operation is shown in the dynamic programming table and how it is possible to transform one word by the best replacement candidate. See Figure 1.

	#	e	x	e	c	u	t	i	o	n
#	0	← 1	← 2	← 3	← 4	← 5	← 6	← 7	← 8	← 9
i	↑ **1**	↖←↑ 2	↖←↑ 3	↖←↑ 4	↖←↑ 5	↖←↑ 6	↖←↑ 7	↖ 6	← 7	← 8
n	↑ 2	↖←↑ **3**	↖←↑ 4	↖←↑ 5	↖←↑ 6	↖←↑ 7	↖←↑ 8	↑ 7	↖←↑ 8	↖ 7
t	↑ 3	↖←↑ 4	↖←↑ **5**	↖←↑ 6	↖←↑ 7	↖←↑ 8	↖ 7	←↑ 8	↖←↑ 9	↑ 8
e	↑ 4	↖ 3	← 4	↖← **5**	← **6**	← 7	←↑ 8	↖←↑ 9	↖←↑ 10	↑ 9
n	↑ 5	↑ 4	↖←↑ 5	↖←↑ 6	↖←↑ 7	↖←↑ **8**	↖←↑ 9	↖←↑ 10	↖←↑ 11	↖↑ 10
t	↑ 6	↑ 5	↖←↑ 6	↖←↑ 7	↖←↑ 8	↖←↑ 9	↖ **8**	← 9	← 10	←↑ 11
i	↑ 7	↑ 6	↖←↑ 7	↖←↑ 8	↖←↑ 9	↖←↑ 10	↑ 9	↖ **8**	← 9	← 10
o	↑ 8	↑ 7	↖←↑ 8	↖←↑ 9	↖←↑ 10	↖←↑ 11	↑ 10	↑ 9	↖ **8**	← 9
n	↑ 9	↑ 8	↖←↑ 9	↖←↑ 10	↖←↑ 11	↖←↑ 12	↑ 11	↑ 10	↑ 9	↖ **8**

Figure 1. Diagram design after Gusfield (1997) [1].

Now, let's go back to the needed iteration number to achieve the aim. As we mentioned in the example i=V=1000000; This value can be limitlessly higher for the Georgian language, though.

Now back to the number of iterations needed to reach the desired goal. As we said, for example i = V = 1000000. However, this value in the case of the Georgian language can be indefinitely high. Based on the results of our earlier studies, we can claim that Georgian possesses words by whose unchanging part (so-called lemma) and the correct concatenation of the matching morphological representatives around 5000 grammatically correct forms of the given word may be generated [8-12]; e. i, **V** in this case will be extremely high, maybe tens of millions. See the picture where using a root „წერ“ and all the word forms have automatically been generated by our program [8]. See Figure 2.

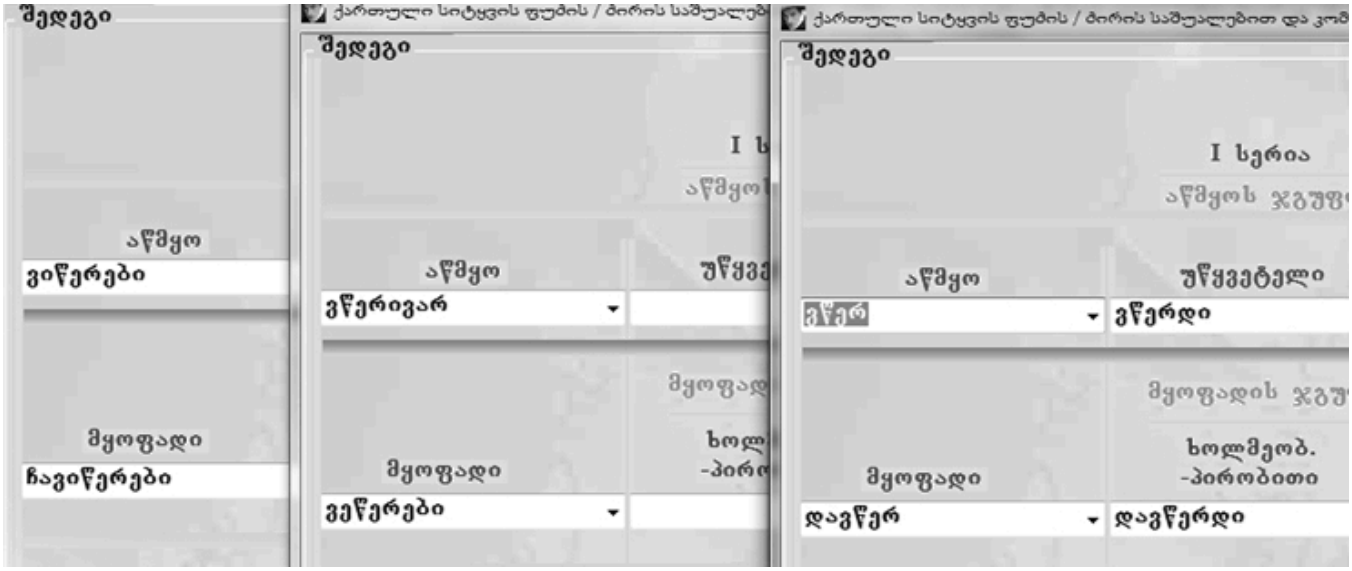

Figure 2. Generated words [4].

ISSN 1512-1232

It presents one side of the issue. On the other hand, since in a decentralized system information is scattered around on different locations, such a high iteration rate for each word will lift the problem to a much higher degree.

Our approach enables to generate all the correct forms automatically by giving the unchanging part of the word. We have already done research in this direction and created an application that, in case of identifying the root by using the necessary morphological atoms, can automatically generate all the grammatically correct forms only of this word. That means, if we say that in the classical case, for example i = V = 1000000, as a result of this optimization, we use only those words that the program itself gave us, i.e. i=V=5000 [8, 13]. The difference is big and obvious. It, also, should be underlined that in the first case V is constant or in worse case a growing value. After the modification, vice versa, V is no longer constant and has a low value in many cases- based on the nature of the word it may not be 5000 but considerably less.

2.2 Multinomial Naïve Bayes Model

Now, let us generalize this method and apply it to other cases, for example, Multinomial Naive Bayes Model (3).

$$P(w_i|c_j)=\frac{count(w_i,c_j)}{\sum_{w\in V} count(w,c_j)} \quad (3)$$

In this case, the probability that the given word or document wi belongs to the given class ci depends on the number of unique words contained in the corpus, i.e., On V. This unit plays a vital role in the quantity of iteration as in the value of probability. In our situation, decreasing V increases the speed and, in addition, the probability - a new value (unit) specific to the context of this algorithm. It is necessary to mention that the classical V - is the unit size and, in our approach- the consequence or collection subset (4).

$$V_{New} = \{v_1, v_2, v_3, \dots, v_n\} \quad (4)$$

In the specific case we have (5):

$$V_{New} = V_{particular} \quad (5)$$

And therefore end up with (6):

$$P(w_i|c_j)=\frac{count(w_i,c_j)}{\sum_{w\in v_{particular}} count(w,c_j)} \quad (6)$$

Now let's consider the general case when the given word is entirely unfamiliar to the model. Without a doubt, according to the formula mentioned above, its quantity will equal zero, and consequently, the classification process in sum will be zero, too. It is obvious that such a case should be avoided.

2.3 Laplace (add-1) smoothing for Naïve Bayes

Hence, let's use the Laplace (add-1) smoothing for Naïve Bayes which is represented in the following formula:

ISSN 1512-1232

$$P(w_j|c)=\frac{count(w_i, c)+1}{\sum_{w\in V} count(w, c)+|V|} \qquad (7)$$

The (6) depicts that even if in the numerator, count (w*i*, c) is equal to zero, it will always have a constant one added to it, which completely excludes the existence of zero in the sum. It should be mentioned that the input of a constant value is normalized-compensated by the number of words in the corpus. Using our method increases the probability and spares the calculating resource expanses, as shown in the formula below:

$$P(w_j|c)=\frac{count(w_i, c)+1}{\sum_{w\in v_{particular}} count(w, c)+|v_{particular}|} \qquad (8)$$

Moreover, our approach considers the cases the implementation of which simplifies the decision-making process for the system. For instance, it implies improving the notion of an error matrix by giving additional mechanisms. However, the latter issue goes beyond the idea of the present paper.

3. Conclusion

In the given paper, we discussed how to reduce the size of the word database and how to optimize existing algorithms. On the other hand, how to combine blockchain system and NLP capabilities with each other. As the example is the State Public Registry of Georgia, it can be said that the implementation of our approach will modify the entire system and can be a trigger for creating an improved language model of Georgian. In addition, it is entirely possible that in this process new visions will emerge due to the peculiarities of the Georgian language. In general, it provides a more flexible intellectual tool.

Reference

[1] Jurafsky, D., Martin, j.h. Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Draft of October 16 (2019).

[2] Mirtskhulava, L., Globa, L., Gulua, N., Meshveliani, N. (2021). Complex Approach in Cryptanalysis of Internet of Things (IoT) Using Blockchain Technology and Lattice-Based Cryptosystem. In: Ilchenko, M., Uryvsky, L., Globa, L. (eds) Advances in Information and Communication Technology and Systems. MCT 2019. Lecture Notes in Networks and Systems, vol 152. Springer, Cham. https://doi.org/10.1007/978-3-030-58359-0_4

[3] L. Mirtskhulava, E. Esiava, N. Gulua, A BLOCKCHAIN-BASED TRUST MODEL FOR BITCOIN CRYPTOCURRENCY AND ITS POPULARITY IN GEORGIA. Computer Sciences and Telecommunications, Issue 2, pp 43-48, 2021

[4] Kardava, I., Gulua, N., Antidze, J., Toklikishvili, B., Kvaratskhelia, T. (2022). Computer Application of Georgian Words. In: Vetulani, Z., Paroubek, P., Kubis, M. (eds) Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2019. Lecture Notes in Computer Science(), vol 13212. Springer, Cham. https://doi.org/10.1007/978-3-031-05328-3_7

[5] Antidze, J. and Gulua, N., Software Tools for Some Natural Language Texts Computer Processing, Computer Technology and Application, vol. 3, no. 3, pp. 219-225, (2012).

ISSN 1512-1232

[6] MelikiShvili, D., The System of Georgian Verbs Conjugation, Tbilisi, Georgia: Logos Press, (2001).

[7] MelikiShvili, D., The Georgian Verb: A Morphosyntactic Analysis, New York, USA: Dunwoody Press, (2008).

[8] Antidze, J. and Gulua, N. Kardava, I., The Software for Composition of Some Natural Languages' Words Lecture Notes on Software Engineering, pp 96-100, (2013).

[9] Kardava, I., Georgian Speechrecognizer in Famous Searching Systems and Management of Software Package by Voice Commands in Georgian Language. Conference Proceedings – Third International Conference on Advances in Computing, Electronics and Communication, pp 6-9, (2016).

[10] Kardava, I. Tadyszak, K. Gulua, N. Jurga, S., The software for automatic creation of the formal grammars used by speech recognition, computer vision, editable text conversion systems, and some new functions. Proceedings Volume 10225, Eighth International Conference on Graphic and Image Processing (ICGIP 2016); 102251Q https://doi.org/10.1117/12.2267687, (2017).

[11] Melikishvili, D., System of Georgian verbs conjugation, Logos Press, Tbilisi, Georgia, (2001).

[12] Melikishvili, D., On Georgian Verb-forms Classification and Qualification Principles, Problems of Linguistics, 1, pp. 30-35 (2008).

[13] Kardava, I., Gulua, N., Toklikishvili, B., Meshveliani, N., Kvaratskhelia, T., Vetulani, Z., Individual Management of Mysql Server Data Protection and Time Intervals between Characters During the Authentication Process. Lecture Notes in Engineering and Comput-er Science, Newswood Limited - International Assotiation of Engineers, pp. 213-217, (2021).

Article received: 2022-06-18