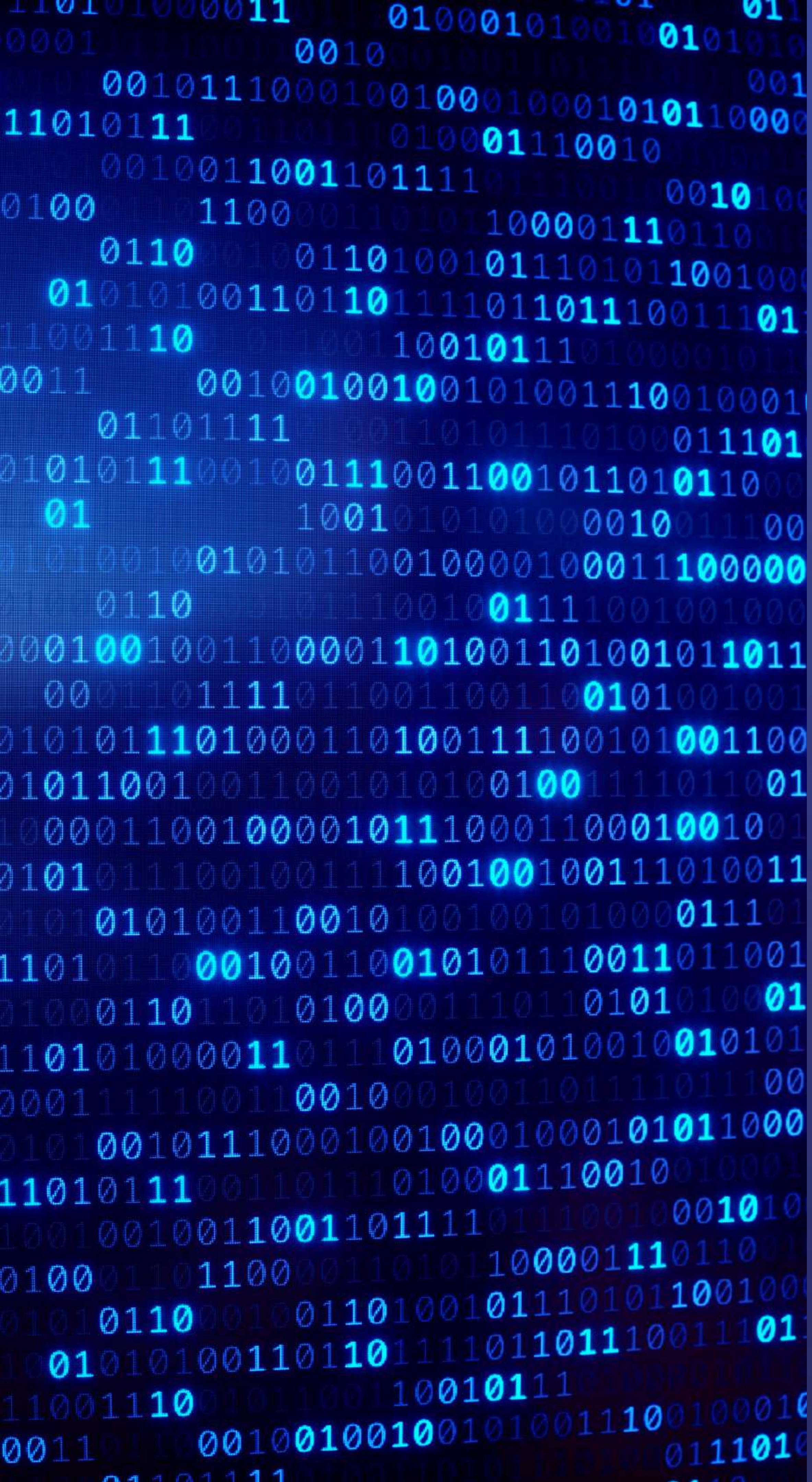


International Conference for Computer Processing
of Turkic Languages "TurkLang"

Using Wikipedia in National Languages as a Dataset ("raw material") for Teaching Neural Network

Shayakhmetova Aigul Rustamovna
International Volunteer Movement Wikimedia
Bashkir Wikipedia volunteer, Ufa

"The Bashkir language is on the brink of dying out"

said Timur Mukhtarov (Candidate of Sociological Sciences), deputy director of Institute of History, Language and Literature, Ufa Federal Research Centre of Russian Academy of Science.
He refers to the evaluation by UNESCO, as the organization assigned a "vulnerable" status to the Bashkir language according to its criteria developed in 2003.

"A language that has no Wikipedia, has no chances"

András Kornai
A Hungarian mathematical linguist

Bashkir section ranks among the first 100 in World Wikipedia Ranking

Bashkir Wikipedia Activists Forum. The volunteers’ many years of service were highly praised by state, departmental and public awards

The attitude towards Wikipedia and Wikimedia related projects

! BASHKIR LANGUAGE MASS MEDIA PROVIDE A REGULAR INFORMATION SUPPORT

? THE ACADEMIC COMMUNITY IS STILL SKEPTICAL ABOUT WIKIPEDIA

This resulted in Bashkir Wikipedia’s popularity growth among Internet users, and volunteer community growth including new volunteers among journalists

WIKIPEDIA
The Free Encyclopedia

We hope that providing a forum for Wikipedia activists in "TurkLang2021" would be a turning point, and Turkic languages linguists will make a tangible contribution to Wikipedia and other Wikimedia projects development in their native languages

Google announced a multilingual variant of BERT project

As a "raw material" they used texts taken from 104 biggest and most active sections of Wikipedia

INCLUDING 9 TURKIC LANGUAGES SECTIONS:

- Azerbaijan
- Bashkir
- Kazakh
- Kyrgyz
- Tatar
- Turkish
- Uzbek
- Chuvash
- South Azerbaijan

This is a breakthrough in human and computer interaction

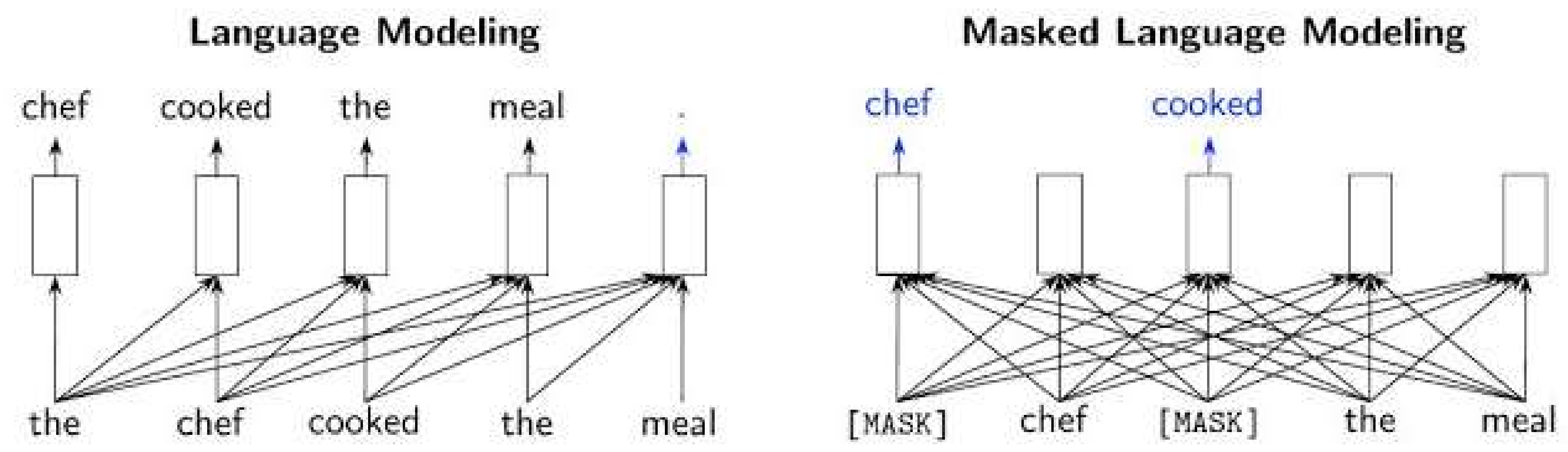
Now a machine can understand a human language,
including slang,
mistakes,
synonyms, and expressions in our speech

We are proud

that thanks to Bashkir volunteers’ activity our mother tongue
became a part of the revolutionary project by Google,
and alongside with other international languages
it advanced froward in the field of Computer Technologies

BERT is a neural network that can learn forms of expression of a human language

The machine can understand the full context of a word, i.e. the terms that precede and follow the word as well as their relations. It "sees" a text in two directions (MLM).
In comparison to other systems that have only one direction (LM).



That is why the model can use large amount of unmarked text data

01 The most expensive stage of pre-training using this model is conducted by Google on the abovementioned languages

02 The results are published

03 Developers can use pre-trained BERT codes and patterns to create their products

04 And this is only the beginning!

In 2021 Google published an article called:

"BigBird: Transformers for Longer Sequences"

One of the advantages of the new model

BigBird can process sequences, which are up to 8 times longer than it was before

Conclusion:

1. Wikipedia is free not only for content consumers, but also for creators of content in their native languages, i.e. there are additional possibilities for native speakers to use their language in everyday life, education, research, art.

2. Wikipedia is not a comprehensive linguistic corpus and was not created for academic and research purposes. This is just a digital corpus of texts in more than 300 languages designed for general public and accessed under Creative Commons license. Nevertheless, world IT companies widely use Wikipedia content for their projects.

3. Wikipedia will not solve all linguistic problems. However, one should not neglect the unique possibilities it provides. Turkic nations need to use every opportunity to preserve and develop their native languages, including generating content for Wikipedia.

Thank you for the attention!

WIKIPEDIA
The Free Encyclopedia