УДК:81' 82+004(045) DOI: 10.35803/1694-5298.2021.1.65-67

САРЫМСАКОВА А.Ж., НИЯЗАЛИЕВА А.Т.,
АЙДАРОВА Л.А.
[1]КГУСТА им. Н. Исанова, Бишкек, Кыргызская Республика

SARYMSAKOVA A.J., NIJASALIEVA A.T.,
AIDAROVA L.A.
[1]KSUCTA n.a. N. Isanov, Bishkek, Kyrgyz Republic
aelita.65@mail.ru, a-nijasalieva@mail.ru, luaza.aidarova09@mail.ru

ИСТОКИ РАЗВИТИЯ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

BACKGROUND OF COMPUTER LINGUISTICS

Макалада компьютердик лингвистиканын башталышы анын багыттары жана компьютердик линвист адистин аткарган иштери жөнүндө маалымат берилди.

Өзөк сөздөр: *лингвистика, компьютердик лингвистика, компьютердик технология, тез которуу, лингвистикалык анализ*

В статье рассматривается история развития, направления компьютерной лингвистики и работа компьютерного лингвиста.

Ключевые слова: *лингвистика, компьютерная лингвистика, компьютерная технология, машинный перевод, лингвистический анализ.*

This article covers the history of development, directions of computer linguistics and the work of computer linguist.

Key words: *linguistics, computer linguistics, computer technology, machine translation, linguistic analysis.*

Mathematical linguistics is a branch of the science of artificial intelligence. It all started in the United States of America in the 1950s. With the invention of the transistor and emergence of new generation of computers along with first programming languages, experiments began with machine translation including Russian scientific journals. In the 1960s, similar studies were done in the USSR (for instance: an article on the translation from Russian into Armenian in the collection “Problems of Cybernetics” of 1964). However, the quality of machine translation was still far inferior to the quality of human translation. The first artificial intelligence systems such as SHGSL were created.

These techniques are outdated, however, popular among students and researchers of academies of sciences dealing with computer linguistics.

From May 15 to May 21, 1958, the first USSR national conference on machine translation was held at the Moscow State Pedagogical Institute. The organizing committee was headed by V. Yu. Rosenzweig and the executive secretary of the organizing committee G. V. Chernov. The full conference program is published in the collection "Machine Translation and Applied Linguistics", vol. 1, 1959 (also known as "Machine Translation Association Bulletin No. 8"). As V. Yu. Rosenzweig recalls, the published collection of conference abstracts ended up in the USA and made a great impression there.

In April 1959, the First USSR Meeting on Mathematical Linguistics was held in Leningrad, convened by the Leningrad University and the Committee for Applied Linguistics. The main organizer of the Meeting was ND Andreev. A number of prominent mathematicians took part in the Meeting, in particular, S. L. Sobolev, L. V. Kantorovich (later - Nobel laureate) and A. A. Markov (the last two took part in the debate). V. Yu. Rosenzweig made a keynote speech “General linguistic

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

theory of translation and mathematical linguistics” on the opening day of the Meeting.

Within the framework of computational linguistics, a relatively new one has also emerged, which has been actively developing since the 1980s – 90s. The study of corpus linguistics, where the general principles of constructing linguistic data corpuses (in particular, text corpuses) are developed using modern computer technologies. Corpuses of texts are collections of specially selected texts from books, magazines, newspapers, etc., transferred to machine algorithm and intended for automatic processing. One of the first corpuses of texts was created for American English at Brown University (the so-called Brown Corpus) in 1962–63 under the direction of W. Francis. In Russia, since the early 2000s at the Institute of the Russian language. V.V. Vinogradov RAS, the National Corpus of the Russian language was developed, consisting of a representative sample of Russian-language texts in the amount of about 100 million tokens. In addition to the actual design of data corpuses, corpus linguistics is engaged in the creation of computer tools (computer programs) designed to extract a variety of information from text corpora. From the point of view of the user, the requirements of representativeness (representativeness), completeness and economy are imposed on the text corpora.

Fields of computational linguistics

• **Natural language processing** (syntactic, morphological, semantic text analysis). This also includes:

1.Corpus linguistics, creation and use of electronic text corpora

2. Creation of electronic dictionaries, thesauri, ontologies.

For example, Lingvo. Dictionaries are used, for example, for automatic translation, spell checking.

3. Automatic translation of texts. Prompt is popular among Russian translators. Google Translate is one of free tools.

4.Automatic extraction of facts from text (extraction of information)
(English fact extraction, text mining)

5.Automatic text summarization.

This feature is included, for example, in Microsoft Word.

6. Building knowledge management systems.

7.Creating question answering systems.

• **Optical character recognition** (English OCR). For example, Fine Reader

• **Automatic speech recognition** (English ASR). There are paid and free software.

• **Automatic speech synthesis**

What does a computer linguist do?

Computational linguistics (referencing paper from English computational linguistics), one of the areas of applied linguistics, in which for the study of the language and modeling the functioning of the language in certain conditions, situations and problem areas, computer programs are developed and used, computer technologies for organizing and processing data. On the other hand, this is the area of application of computer language models in linguistics and related disciplines. As a special scientific area of computational linguistics, it took shape in European studies in the 1960s. Since the English adjective computational can also be translated as "computational", the term "computational linguistics" is also found in the literature, but in domestic science it acquires a narrower meaning, approaching the concept of "quantitative linguistics".

Computer linguists are engaged in the development of algorithms for recognition of text and sounding speech, the synthesis of artificial speech, the creation of semantic translation systems and the development of artificial intelligence itself (in the classical sense of the word - as a replacement for human - it Speech recognition algorithms will be used more and more in everyday life - “smart homes” and electronic devices will not have remotes and buttons, and instead will use a voice interface. This technology is being refined, but there are still many challenges: it is difficult for a computer to recognize human speech, because different people speak very differently. Therefore, as a rule, recognition systems work well either when they are trained for one speaker and are already adjusted to his pronunciation features, or when the number of phrases that the system can recognize is limited (as, for example: in voice commands for the TV is unlikely to ever appear, but various expert systems based on data analysis).

The specialists in creating semantic translation programs still have a lot of work ahead: at the moment, good algorithms have been developed only for translation into and from English. There are many problems here - different languages are semantically arranged differently, this differs even at the level of phrase construction, and not all meanings of one language can be conveyed using the semantic apparatus of another. In addition, the program must distinguish between homonyms, correctly recognize parts of speech, and select the correct meaning of a polysemantic word that suitsthe context.

Synthesis of artificial speech (for example, for home robots) is also painstaking work. It is difficult to make the artificially created speech sound natural to the human ear, because there are millions of nuances that we do not pay attention to, but without which everything is not "right" - false starts, pauses, hitching, etc. The speech flow is continuous and at the same time discrete: we speak without pause between words, but it is not difficult for us to understand where one word ends and another begins, and for a machine this will be a big problem.

The largest direction in computational linguistics is associated with Big Data. After all, there are huge corpuses of texts like news feeds, from which you need to extract certain information - for example, highlight news feeds or sharpen RSS to the tastes of a certain user. Such technologies already exist now and will develop further, because the computing power is growing rapidly. Linguistic analysis of texts is also used to ensure security on the Internet, to search for the necessary information for special services.

Where to study to become a computer linguist? Unfortunately, specialties related to classical linguistics, and programming, statistics, data analysis, are quite strongly separated. And, in order to become a digital linguist, you need to understand both. Foreign universities have higher education programs in computational linguistics, while our best option is to get a basic linguistic education, and then master the basics of IT.

It's good that there are many different online courses now. Now IT companies are actively trying to interact with higher education institutions.

Computational linguistics demonstrates quite tangible results in various applications for the automatic processing of texts per language unit. Its further development depends both on the emergence of new applications and on the independent development of various language models, in which many problems have not yet been solved. The most elaborated are the models of morphological analysis and synthesis. The syntax models have not yet been brought to the level of stable and efficient working modules, despite the large number of proposed formalisms and methods. Even less studied and formalized are models of the level of semantics and pragmatics, although automatic processing of discourse is already required in a number of applications.

It is worth noting that the already existing tools of computational linguistics itself, the use of machine learning and text corpora, can significantly advance the solution of problems.

References

1. Gorodetsky B.Yu. Computational linguistics: language communication modeling. - New in foreign linguistics. Issue XXIV, Computational Linguistics. M., 1989

2. Baranov A.N. Introduction to Applied Linguistics. M., 2000

3. Popov E.V. Communication with computers in natural language. M., 1982

4. Sadur V.G. Speech communication with electronic computers and the problems of theirdevelopment. - In the book: Speech communication: problems and prospects. M., 1983

5. Subbotin M.M. Hypertext. A new form of written communication. - VINITI, Ser.Informatics, 1994, vol. 18