GALAXY INTERNATIONAL INTERDISCIPLINARY RESEARCH JOURNAL (GIIRJ)
ISSN (E): 2347-6915
Vol. 11, Issue 05, May (2023)

LEXICAL MEANS OF SUBSTANTIAL CORRESPONDENCE, SIMILARITY AND FORMAL DIFFERENCE IN SIMPLE SENTENCES AND THEIR LINGUISTIC MODELS

Muratbekova Shoira Dilshodbek kizi
Tashkent State Named After Alisher Navoi University of Uzbek Language and Literature
Specialization in Computer Linguistics 2nd Year Graduate Student

ABSTRACT

The development of research in Uzbekistan, the emergence of original scientific texts requires the development of a certain linguistic control (to prevent duplication). For this, it is necessary to create software that determines the level of similarity of the content of scientific texts available in electronic form.

Keywords: model, linguistic model, Levenstein distance, Shingle's algorithm, Jaccard's algorithm

Also in linguistics, as a model concept, a linguist's reordering of language in artificial vision, i.e. bringing it to a more simplified form than its usual appearance. In this place, for linguistic purposes, it imitates aspects, words, speech patterns (repeats behavior) related to the original version of the language.

There are many definitions of models in linguistics:

model - type, pattern (language pattern) of any text units (words, sentences);

model - symbols, schemes for describing language objects;

model is a formalized theory of a structure with a fixed metatext.

The main goal of modeling in linguistics is the integration of the individual modeling of linguistic ability and thereby creating conditions for quick, easy and convenient implementation of his work. The concept of a linguistic model originated in structural linguistics, but came into scientific use in the 60s and 70s. In modern linguistics, the meaning of the term "model" was mainly put forward by Yelmslev[1] under the name of the term "theory". The model is considered to be worthy of the name of theory only if it is expressed clearly enough and formalized enough [1].

Each model should ideally be implemented on a computer.

In linguistics, the concept of "linguistic model" appeared as a result of the influence of structural linguistics and later led to the emergence of computational linguistics [2]. It turns out that the linguistic model as a phenomenon connects structural linguistics and computer linguistics. Building a model is not only a means of reflecting linguistic phenomena, but also an objective practical criterion for verifying the truth of language knowledge.

It is no exaggeration to say that modeling, together with other methods of language learning, works as a means of deepening knowledge about the hidden mechanisms of speech activity, its transition from relatively primitive models to more meaningful models that more fully reveal the essence of language.In linguistics, modeling has its principle as a system, and some of its subsystems model others, for example, the system of written speech is a model of spoken speech; in written language we work with several models (printed, handwritten);

expression plan content plan model.

The method of modeling is usually based on sign systems, but language itself is also a system of signs, that is, modeling words with words. Any model, including a linguistic model, must be formal.

If the model can clearly and concisely show the initial information and texts and the rules for working with them (rules for the formation or placement of new objects and thoughts), then the intended goal will certainly be formed and formality will appear in this form.

Will come.

Ideally, any formal model is a mathematical system.

Therefore, in a certain sense, the concept of formalism is required to be equivalent to the concept of mathematics, precision or uncertainty. Formality, precision, ambiguity - these are the characteristics of the language in which the theory is presented. By itself, this feature does not ensure that the predictions of the official theory match objective experimental data. The correctness of a theory makes it possible to establish specific experiments capable of confirming or refuting it, but it is logically wrong to conclude without practical experiments that there is a necessary logical connection between the accuracy of the theory and its reality.

Event that happened.

A formal model is linked to experimental data through one interpretation or another. Model interpretation is characterized by the fact that, instead of model objects (symbols), objects of a certain subject area are represented, for example, probabilistic or fixed rules of language substitution.In the first parts of our work, we noted the current situation related to copying and its prevention, that is, although the legal basis for copyright protection is sufficient, scientific approaches in this regard are far behind. In philosophy, it is considered a law that the change in quantity affects the change in quality, despite the independence of the Republic of Uzbekistan more than thirty years ago, there is no solution to the urgent problems that can compete with world linguistics in the issue of conducting research, especially in linguistics.

Of course, the law is strong, but according to the view that human needs are stronger than it, the illegal appropriation of other people's work and the avoidance of scientific ethics remain a serious problem.

As we mentioned above, the legislation can determine what can be done, what can't be done, and what can be punished when caught doing a prohibited act. However, if the violation of copyright is not scientifically substantiated, if one does not have methods and means of deep analysis, it will be inappropriate and impossible to expect an effective result. Today, it is a serious problem to determine the original text of scientific and research works carried out in developed countries. Posting scientific works on the Internet by higher education institutions, finding such works cheaply and easily using the Internet, is the reason for the escalation of problems in this regard. Of course, we mentioned above that there are many commercial programs abroad that check the originality of scientific texts.

However, the fact that there are more commercial offers, in particular, the continuous development of "smart synonymizers" and methods to cheat the program, is a serious and urgent problem.

Using online anti-plagiarism programs on the Internet, we have developed our own proposals and recommendations for preparing a linguistic base for an anti-plagiarism program that determines the level of semantic similarity of simple sentences in scientific texts in the Uzbek

language and determining its working methods.In computational linguistics, the term linguistic module is gaining importance today.

Therefore, the transfer of natural language to computer language, that is, the discovery of ways of text processing with the help of a computer system, is observed.

In this regard, linguistic programs of foreign languages have been developed and are being improved today. A linguistic module is an independent component of such linguistic programs, that is, a part of the software that covers a specific linguistic process [2]. In fact, the theory of language originates from the characteristics of the existing language, it is the preparation and processing of its specific aspects for systematic use according to a certain order.

In other words, theory came from practice, and today the process of perfecting the theory and returning it to practice is much more active.

The Jaccard measure (floristic commonality coefficient, French coefficient of communication, German scientist)[3] is used to determine similarity in texts (proposed by Paul Jacquard in 1901 is a binary similarity measure).[1] : KJ = ca + b - c {\displaystyle K_ {J} = {\frac {c} {a + bc}}} K_ {J} = {\frac {c} {a + bc}}, where a

- the number of species in the first test area, b - the number of species in the second test area, c - the number of common species for sites 1 and 2.

This is the first known level of similarity.

In the literature, the last name of the author of the coefficient is also given as Jacquard. The Jaccard coefficient is actively used in ecology, geobotany, molecular biology, bioinformatics, genomics, proteomics, informatics and other fields in various modifications and notations.

The Jacquard measure is equivalent to the Sorensen measure and the Sokal-Snit measure for finite sets.In the Uzbek language, one sentence can be expressed in different ways, for example, by changing the place of words or by replacing words with synonyms. The need to determine the similarity of two sentences arose when solving a small practical problem. Simple measurements and aggregated statistics are used to determine the coefficients.

Briefly, the task can be expressed as follows: "New sentences come with a certain frequency in different sources. The output should be filtered so that there are no two sentences about the same fact."

Comparing two sentences

There are several ways to solve the problem of determining the degree of similarity between two strings.

Levenshtein distance[4]

Returns a number indicating how many operations (add, delete, or replace) to perform to convert one string to another.

Features:

simple program;

depends on word order;

output is a number;

the output must be compared to something.

For example, "I teach at school."

and "I teach at a university."

The more he changes these sentences to fit each other, the lower the percentage, the less the change, the higher the percentage.
That is, "I" did not change, "university" did not change, "class" did not change, "I give" did not change. In this case, 75 percent are transferred.
However, the two examples of the sentences differ sharply in terms of their meaning (a teacher at a university and education at a university have many aspects that differ from a school and a school teacher).
"I teach at school."
and "I teach at school" does not match the two sentences at all and remains 100% copied by itself.
Shingle Algorithm[4]
It divides texts into shingles (in English - scales), that is, chains of 10 words (with intersections), applies hash functions to shingles, takes matrices, compares them with each other.
Features:
to implement the algorithm, it is necessary to study the mathematical part in detail;
works on large texts;
does not depend on the order of the sentences.In this algorithm, lines and words are compared, the change of word order is not important, and the percentage is calculated based on the number of redundant words.
"I'm a teacher at school."
and "I am a teacher at school."
In this case, he does not know that "teacher" and "teacher" are synonyms, and 70% of them think that they are similar.
Jacquard algorithm
In this algorithm, letters are compared, not words.
"I am a teacher at school."
and "I'm a school teacher."
In this place, the additions -da and -si are redundant, 76% are similar. As another example of this, we acknowledge that there are serious shortcomings by pointing out that the words asr and asir are 90% similar.When the components of logical vectors are used, that is, components that take only two values 0 and 1, the measure is known as the Tanimoto coefficient or the augmented Jaccard coefficient.
If objects are compared with the occurrence of species (probability interpretation), that is, if the probability of meeting is taken into account, then the analogue of the Jaccard measure is the Iversen probability measure [4].In this method, it is possible to check only compatibility, that is, when words are changed to synonyms, when fragments that are not syntactically related to the sentence are added, difficulties arise in determining semantic similarity.
That is, only words that are exactly similar are compared in this method.
And for us, it is important to determine the semantic similarity of words that are not exactly similar.
In conclusion, it can be said that for anti-plagiarism programs that determine the degree of similarity of scientific texts, it is necessary to create a special list of fragments that do not enter into a syntactic relationship with the sentence. It is appropriate to use explanatory and

synonym dictionaries of the Uzbek language to model the state of synonymy of words and phrases in phrases when determining the degree of semantic similarity of simple sentences in scientific texts [5].

LIST OF USED LITERATURE

1. https://studopedia.info/2-73325.html
2. Abdurahmonova N. Computer Linguistics. - Tashkent: Nodirabegim,2021. 317b.
3. Abjalova M. Modules of editing and analysis programs. - Tashkent: Nodirabegim, 2020. - 16p.
4. https://habr.com/ru/post/341148/
5. Murtazayev A. Content similarity and linguistic models of simple sentences in scientific texts. - Kokand - 2021. - 133b.