STEPS IN THE PROCESS OF IDENTIFYING FULL-TEXT SIMILARITIES IN A MULTILINGUAL DATABASE

Avezmatov Ikhtiyor Davlatyarovich

Urganch State University, ixtiyoravezmatov07@gmail.com

Shermatov Boburjon Inoyat O‘g‘li

Urganch State University, bobur shermatov912@gmail.com

Babayev Saidmukhammadjon Saidkamolovich,

Urganch State University, saidmuhammadbabayev@gmail.com

Bekchanov Hikmat Mansurbek o’g’li

Urganch state university, bekchanovhikmat209@gmail.com.

Abstract : The article describes the processes of identifying similar texts from a multilingual full-text database: changing the text format, morphological analysis of words in the text, and storing them in the database.

Keywords: Database, morphological analysis of words, text comparison.

INTRODUCTION

Systems designed for text processing are divided into anti-plagiarism detection systems, text segmentation systems, systems designed for translation, systems designed for information retrieval, etc. These systems, which belong to the field of automatic text processing (ATP), each have their own specific goals and objectives. Some of the tasks of the systems are similar to each other, while others are fundamentally different. For example, let's look at systems designed for text translation and information retrieval. In systems designed for text translation, the main focus is on segmenting the text into parts, morphologically analyzing them, and translating the analytical results based on a

dictionary [1, 2]. In systems designed for information retrieval, the main focus is on comparing terms.

of texts also belongs to MAIB and has a much more complex structure than the processes performed in the above-mentioned systems. Because, in order to identify multilingual similarities of texts, it is necessary to solve the problems of comparing them with each other in addition to translating the texts using a machine [3, 4]. Currently, there are a number of developments in this area, which mainly consist of: text analysis, as well as storage in a database and information retrieval from a full-text database.

Table 1 .

Parts of the CLAD method

o.	Naming	Brief task
.	Normalize text formatting .	Converting text from various formats (.doc, .docx, .pdf, .html, etc.) to plain text.
.	Text words package to the shape transfer and meaning giver words to determine .	Simple the text regular actions based on into parts separation , and harvest was terms from the collection comparison to the process chaos to do general meaningful words take throw
3.	The word technician core part to determine .	In the sentence the word technician in terms of unchanging part to determine . This The morphological rules of each language are taken into account in the process .
.	Terms one from the language other to the language translation to do	Text car using translation to do
.	Lexical analysis of terms.	Various shaped meaning one kind terms to determine .

.	Saving terms in DB.	Texts terms set in the form of non-relational In MB save .
.	Texts each other similarity check .	DB data from base T to text similar texts to determine .

Converting data to plain text . As is known, there are currently various formats for describing electronic resources, which are fundamentally different from each other in terms of storing and displaying data. Among them, .DOC (.DOCX), .PDF, .HTML formats are widespread, and currently there are various models, algorithms and software tools that convert data in these formats to plain text. As one of them: we can cite the Apache Tika system. This system is not only free, but also open source. This allows you to change the software part of this system if necessary.

texts into words and extracting meaningful words from them. When dividing multilingual texts into words, the speech is divided into parts based on the punctuation marks used to divide words according to the rules of the spoken language (“-”, “,”,”:”,”;”,,”.”). Various methods can be used to divide text into words. The easiest of these is the method based on regular expressions. Below is a regular expression designed to divide text in the Latin alphabet into words.

~ [A - Z]. * [. , : ! ? ;] (? = \s|$) ~ s

Based on this regular expression, we can split a text described in the Latin alphabet into words using any programming language.

The process of extracting meaningful words from the text is carried out by removing words that do not have any meaning in the natural language (for the Uzbek language: words belonging to the auxiliary word class and pronouns) from the resulting set of words.

Of words technician core part definition . As you know, the text in the content words every kind additions with comes . This the situation Turkish , Slavic in their

languages in texts many our observation possible . Machine using words compared to additions various was two word to each other equal It will not happen . This is texts similarity in inspection the result wrong to be take This stage is currently carried out using two methods: using a dictionary and using algorithms developed based on the morphological rules of the language. Based on the morphological rules of the language, there is a Snowball system for determining the stem of a word in a text, which consists of more than 20 algorithms designed for natural languages.

Translation of words . As is known, in the process of comparing multilingual data, initially, the two texts being compared are brought into one language. If the texts in the database are in different languages, the compared text will have to be translated into the natural language of the database each time. This slows down the process of checking the similarity of the data. To optimize this process, we select the main one from among the natural languages in which the L_m data entered into the database is written $L_1, L_2, .., L_n$. Each piece of data entered into the database if it L_m if not written in the language, initially L_m is translated into the language, then stored in the database.

At this point, let's revisit the steps in the dictionary-based machine translation process. The following figure shows the steps in this process.

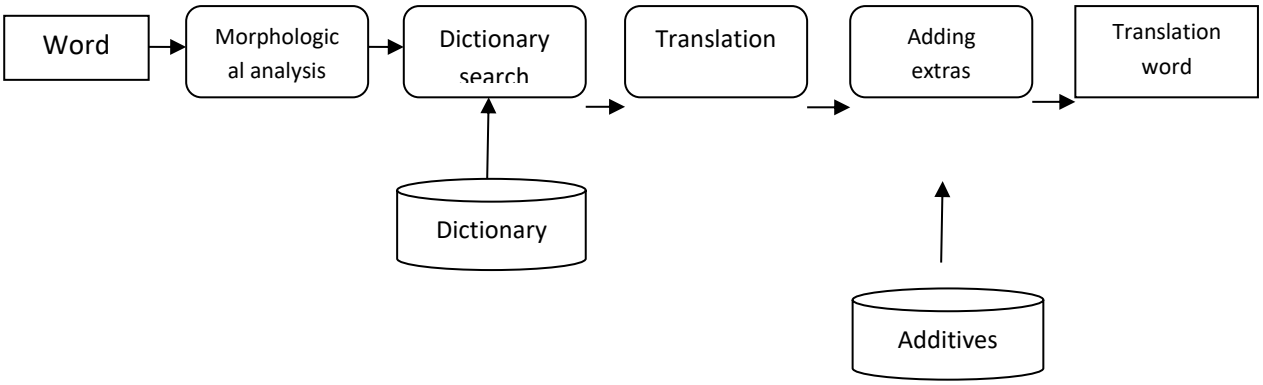

Figure 1. Stages of the translation process using a dictionary

As is known, the main requirement for the process of machine translation of text is the semantically correct selection of translation forms of words in a sentence. For example, let's take the English word "face". This word has such translation forms in Uzbek as "yuz", "chehra", "aft", "bashara", "sath", "usti". If we are talking about a face, then the translation forms "yuz", "chehra", "aft", "bashara" can be used, and if we are talking about the upper

part of something, then the translation can be done in the form "sath", "usti". In the translation process, depending on the meaning of the sentence, one of several translation forms can be used. In the process of determining the multilingual similarity of texts, all forms of translation must be considered.

Comparison of texts. In the process of comparing multilingual texts, texts are described in the form of sets of words. After the given texts D and T are respectively reduced to the form of sets of words, their similarity is determined based on the sets of words that are the same in the content of the texts.

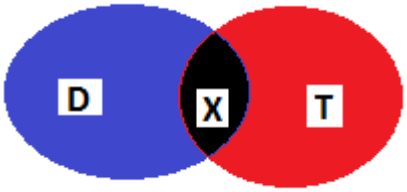

Figure 2: Similar parts of the texts

The similarity of texts depends on the number of common words in their content, and the more similar words they have, the more similar the texts are. If the texts are described in different languages, then initially the words in the text must be translated using a dictionary and the set of words must be brought into the same natural language.

In Figure 2 above, the common parts of texts D and T, described as sets of words, are marked with X. The larger the size of the part X, the more similar the texts are to each other, and vice versa, the smaller the size, the more different the texts are from each other.

Areas of application. The method of determining multilingual similarity of texts is of great importance in solving current problems in many areas. Some of them are listed below:

- In automated information library systems - in these systems, searching for information from full texts linked to a bibliographic record, and identifying a set of bibliographic records that match it based on the abstract text;

- In electronic archives – identifying similar sources among large volumes of archival data;
- In electronic document exchange systems - searching for information from sources in various formats (Word, PDF, Excel);
- In anti-plagiarism systems - in the process of detecting plagiarized documents in copied, hidden, or translated forms.

In conclusion, it can be said that, based on algorithms for determining the similarity of texts, and their application to information and library systems, it is possible to search for information among full-text information in an electronic catalog, and to automatically form an electronic catalog based on abstract texts.

References

1. Gipps, Bela; Meuschke, Norman (September 2011), "Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence", Proceedings of the 11th ACM Symposium on Document Engineering (DocEng2011) (PDF), ACM, pp. 249–258, doi:10.1145/2034691.2034741, ISBN 978-1-4503-0863-2

2. Bermudez S.X. On the method of identifying significant textual passages as a basis for textual comparison // Informatizatsiya i i svyaz. – 2016. – No. 3. – C. 231-219.

3. Atadjanov J.A. — Models of Morphological Analysis of Uzbek Words // Kibernetika i programmirovanie. – 2016. – No. 6. – S. 70 - 73. DOI: 10.7256/2306-4196.2016.6.20945. (05.00.00; #45).

4. Atadzhanov J.A. Organization of the process of searching for full-text information in electronic libraries and archives // Scientific and methodological journal "Kutubkhona.uz". No. 4(44) 2019.