Hybrid language processing approach

Havrashenko Anton[1]

Barkovska Olesia[2]

[1]*Kharkiv National University of Radio Electronics, 14 Nauky Ave, Kharkiv UA-61166, Ukraine, anton.havrashenko@nure.ua*

[2]*Kharkiv National University of Radio Electronics, 14 Nauky Ave, Kharkiv UA-61166, Ukraine, olesia.barkovska@nure.ua*

Abstract. *Creating a computer system for translating text from hybrid languages into any arbitrary language is a challenge not only for our country. The global trend of borrowing words from other languages is spreading rapidly and leading to the emergence of new so-called "mixed" languages. We propose an algorithm for constructing a dictionary and presents translation results that show up to 65% translation accuracy.*

Keywords: *translator, programming language, internet, dictionary, text processing.*

I. Introduction and Problem statement

In the modern world, there are a large number of different languages that can be mixed, to borrow foreign words, leading to the emergence of so-called hybrid languages and slang. Slang - very informal language that is usually spoken rather than written, used especially by particular groups of people[1]. Online translators like Google, Yandex can translate only around 100 literary languages. In addition, adding new words for translation is usually impossible, which is not enough in the modern world.

Today, there are no specialized translators available for translating slang or hybrid languages, but their number is constantly increasing. Previously, new slangs were born during the influence of neighboring countries upon them and depended on a particular territory or among people of a particular profession. But they were quite similar to the original, with the addition or change of not usually a large number of words. Nowadays, slang is formed under the influence of a certain book, movie, game and may not be understood by a stranger.

This causes the relevance of the algorithm for constructing the dictionary to translate text from hybrid languages and slang into existing languages of the world, and also developing a "mixed language translator" application based on the proposed algorithm.

II. Aims and tasls of the work

The main purpose of the project is to implement processing of the input text, turning it into a dictionary and the opportunity to use this dictionary to translate.

To achieve this goal, the following tasks must be solved:
- create an algorithm for constructing dictionary to translate text from hybrid languages and slang existing world languages;
- developing a "mixed language translator" application based on the proposed algorithm.

III. Problem solution and Results

To accomplish these tasks, a software package of 10 programs was developed:

№1 - Initial Text Processing; №2 - bilateral improve processing dictionary; №3 - improving translation of similar languages by the n-gram algorithm;[2] №4 - improving translation of similar languages by Knut-Morris-Pratt algorithm;[3][4] №5 - search phrases, phraseology and translation for them; №6 - a combination of 2 words in one language[5] №7 - build vocabularies in other languages and improve vocabularies using already built ones; №8 - supplementing the finished dictionary with other texts; №9 - splitting the dictionary into dictionaries with a small percentage of text intersection; №10 - translator.
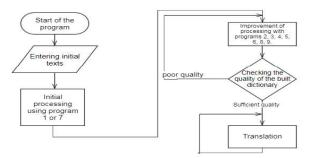

Figure 1. Flowchart of the algorithm

The algorithm have the following structure. Programs 1 and 7 are input and are only used for initial word processing. All programs except 1, 7, and 10 improve the dictionary until the required quality of the dictionary is reached. After achieving the required quality, the dictionary is transmitted to the translator program, which interacts with the user translating the input data.

IV. Conclusions

There is potential for further work on the system. During the testing, no critical errors were detected at any of the stages of operation.

For further development, it is suggested to consider the problem of changing the translation by re-entering the same data and entering incorrect data. Also possible to develop the idea of common vocabulary that is stored on a server and the translation via the Internet.

References

[1] Cambridge Dictionary [Electronic resource] URL: https://dictionary.cambridge.org/dictionary/english/slang

[2] N-gram [Electronic resource] URL: https://uk.wikipedia.org/wiki/N-грам

[3] Prefix function. Knut-Morris-Pratt algorithm [Electronic resource] URL:https://e-maxx.ru/algo/prefix_function

[4] S. Aygün, E. O. Güneş and L. Kouhalvandi, "Python based parallel application of Knuth-Morris-Pratt algorithm," 2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), Vilnius, 2016, pp. 1-5.

[5] Kevin McTait, Arturo Trujillo. A Language-Neutral Sparse-Data Algorithm for Extracting Translation Patterns. Proceedings of 8th International Conference on Theoretical and Methodological Issues in Machine Translation, August 1999, Chester, UK