PARALLEL CORPUS IN THE PRACTICE OF TRANSLATING SOCIAL TEXTS

Ass. professor Zamira Abdullaeva Shamshaddinovna
Ass. professor Marhamat Haydarova Yunusovna
PhD student Guzal Shikhnazarova Alisherovna
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

Annotation

In this article, corpus linguistics, corpus and its parallel corpus, a word about ideas, its structure, corpus types, token, lemma, stemming. Today, the theoretical and practical significance of the corpus is Uzbek in linguistics, in a wider study of the existing possibilities of the language,

in linguistics identifying problematic aspects, creating electronic dictionaries, learning the language in improving the efficiency of using modern information technologies, in the language in dealing with issues such as automatic translation, search and computer analysis, of language shows that there is a need to build a corps in special fields.

Key words: corpus, corpus linguistics, parallel corpus, translation corpus.

Natural language processing is a rather complex and complex direction, where several stages of data processing can be observed. One of the stages is morphological analysis, which reveals such properties of the word as "root", "stem" and "affix(es)". It should be noted that the combination of possible analyzes of the word form may vary depending on the number of combinations of the word. In agglutinative languages, word forms are built by concatenation of affixes, and this indicates the possibility of changing the meaning of a word with one morpheme. Turkic languages are one of such languages [1]. The analyzer proposed in the framework of this work is designed for morphological analysis of the word form of the Turkic family of languages. Among the methods used to develop a morphological analyzer, word stemming stands out [2]. This algorithm is quite simple in concept and directly in implementation. The simplicity lies in the fact that during stemming, suffixes are truncated until we reach the root of the word [3]. But, as a rule, stemming serves only as the basis for such analyzers, since if implemented incorrectly, an erroneous result can be observed. For example, stemming can truncate more than necessary, and as a result we can see a set of letters instead of a word root. To exclude such cases, it is necessary to add additional methods for processing word forms. In further sections, we will talk about the proposed solution for the morphological analysis of word forms in Turkic languages.

In corpus linguistics at the beginning of the new millennium corpora of English as an international language dominate, they are mainly created, studied and used in teaching and lexicography. For example, the Corpus of Contemporary American English (COCA https://www.english-corpora.org/coca/), British National Corpus (BNC http://www.natcorp.ox.ac.uk/), Just the Word (http://www.just-the-word.com/), Phrases in English (http://phrasesinenglish.org/), Open American National Corpus (http://www.anc.org/), Cambridge English Corpus, Michigan Corpus of Academic Spoken English (MICASE https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple), International Corpus of English (http://ice-corpora.net/ice/index.html), Word Neighbors (http://wordneighbors.ust.hk/), Corpus of Historical American English (COHA https://www.english-corpora.org/coha/), Collins Corpus (Bank of English https://collins.co.uk/pages/elt-cobuild-reference-the-collins-corpus) and others[1]. In modern corpus linguistics, there are two types of parallel corpus: 1) multilingual corpus (Comparable/Multilingual Corpora); 2) translation corpus (Translation Corpora).

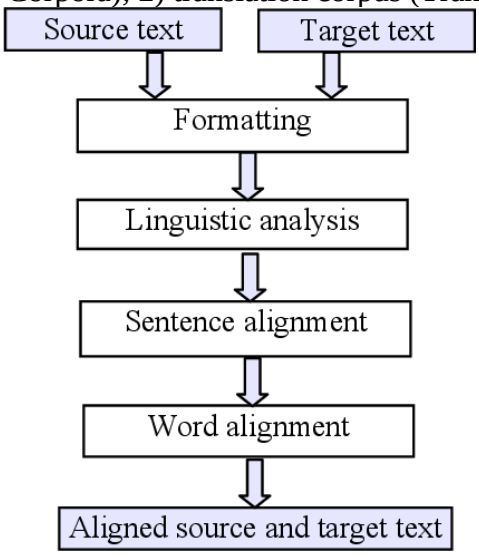

Figure 1. Modules of Uplug

The structural composition of the body with this characteristic is based on its purpose can be different[2]: 1) in the usual text form referring to the translation;

2) in the form of "texts in the mirror" convenient for comparison; 3) in the form of a database.

According to the structure and arrangement of texts, adaptation of units There are several types of parallel case:

1) one-way (text translated from English to Russian)

2) two-way (text translated from English to Russian and from Russian into English (back translation))

3) multi-directional (translated from English to Russian, German, French text)

Parallel corpus is the structural structure of the translated language, phrases and words serves to monitor its possibilities in the context. Such cases are as follows

used for purposes [3]: 1) in comparative linguistics (in order to analyze two language structures); 2) in the field of translation [4] (search for the equivalent of the original text in other languages purpose); 3) in the field of automatic translation; 4) in language didactics; 5) in lexicography.

The text and its translation [5] are another part of the accompanying electronic collection There are also views. Bitext program is one of such tools. Parallel text (bitext) is a set of texts with translation in another language.

Bitexts is an "alignment tool" or "for bitext special computer known as bitext tool

through programs. These programs are original and The content of the translated text is mainly simple, with different syntactic units adapts it in the form of a sentence. The bitext database or It is also called a bilingual corpus, which allows to display different connections acts as a database (directory).

Parallel case: first and last view. The field of corpus linguistics is foreign for language teaching methodology and translation, computer linguistics is gaining practical importance [6]. Parallel in foreign language and translation theory education M. Barlow, McKenery, Baker, Zanneti, Arenberg, The opinions of scientists such as Blank, Brown, Church, Gale, Davis, Foster are valuable.

The idea of bitext belongs to Brian Harris, who originally bitext in 1988 developed the concept; This concept was later adopted by the University of Montreal developed by group scientists; RALI (Recherche appliqué en linguistique informatique or Applied Research in Computational Linguistics//Prikladnye issledovaniya v vychislitelnoy lingvistike» activity was carried out in the association. Linguists and programmers work in the group conducted; studied text processing problems. Bitext for the first time the concept was realized by Pierre Isabelle and Claude Bedard. "Corpus of parallel texts for scientific and practical purposes (including foreign for the purpose of teaching languages) is being implemented.

Concusion

Creating a parallel corpus involves several steps: textalignment, text marking, search interface design. The editing process is primarily original in translation is used to ensure that there is a fragment that matches the copy. After that, the same parts of parallel texts with each other is compared. The question that arises at the initial stage is what should be leveled. You you can align the word with the word, but this is often for a number of reasons It turns out to be almost impossible: a collection of tokens, stable in different languages phrases do not match. Also, texts are aligned by sentences, but that's it problems may arise: the number of sentences or paragraphs is also different can do. The following original Uzbek text has been translated into Turkish we can see the translation. Aligning words with words is a problem found its proof. Number of words in Uzbek language with Turkish translation We explain that it is not compatible with this sentence. 13 in the example given in Uzbek the word was involved. There are 9 words in the Turkish translation

References

1. Маник, С. А. (2019). Параллельный корпус в практике перевода общественно-политических текстов (с английского на русский). *Современные исследования социальных проблем*, *11*(4), 225-245.

2. Sobirova, N. G. Q., Akramajanova, M., & Sobirov, S. G. O. (2022). PARALLEL KORPUS XUSUSIYATI, TUZILISHI, INTERFEYSI. *Academic research in educational sciences*, *3*(5), 819-824.

3. Abdurakhmonova, N., Alisher, I., & Sayfulleyeva, R. (2022, September). MorphUz: Morphological Analyzer for the Uzbek Language. In 2022 7th International Conference on Computer Science and Engineering (UBMK) (pp. 61-66). IEEE.

4. Shixnazarova, G. A., & Gulyamova, D. R. (2016). Modeling of the process of examination data mining techniques in the study of English language. Информатика: проблемы, методология, технологии, 900.

5. Abdullayeva, Z. S., & Shihnazarova, G. A. (2023, April). INTELLEKTUAL TAHLIL ASOSIDA CHET TILI BILIMINI TEKSHIRISHNI MODELLASHTIRISH. In *INTERNATIONAL SCIENTIFIC CONFERENCES WITH HIGHER EDUCATIONAL INSTITUTIONS* (Vol. 1, No. 14.04, pp. 150-152).

6. Каримова, В. А., & Шихназарова, Г. А. (2015). Инглиз тилини ўрганишда билимларни текшириш жараёнини моделлаштириш. *Современное образование (Узбекистан)*, (5), 19-23.

7. Alisherovna, S. G. (2024). MATHEMATICAL MODELING OF MACHINE LEARNING ALGORITHMS. *«CONTEMPORARY TECHNOLOGIES OF COMPUTATIONAL LINGUISTICS»*, *2*(22.04), 418-420.