POST-EDITING OF WORDS IN KAZAKH SENTENCES FOR INFORMATION RETRIEVAL

[1] **SHORMAKOVA A.,** [2]**ZHUMANOV ZH.,** [3]**RAKHIMOVA D.**

[1,2,3] Al-Farabi Kazakh National University, Kazakhstan,Almaty

E-mail: [1]shormakovaassem@gmail.com, [2]z.zhake@gmail.com, [3]di.diva@mail.ru

ABSTRACT

In this paper, the proposed method for determining incorrect words for any context is described in more details. Also, the paper describes the completeness of the meaning of words in Kazakh sentences for incorrect words in post-editing. And the second main task is to find the most suitable variant for these words found by the full meaning of the word for any context. According to the results of the study, the relevance and weighty significance for the Kazakh language was shown to determine the full meaning of the word. After the work done, this methodology shows that it is very relevant and effective especially for the Turkic languages. Although the system is aimed to find the wrong words and analyze these words for post-editing, but these modules can be considered as separate modules for working with contexts. The work was also done for information retrieval and for the analysis of the initial input data. In this work, the essence of the work of a search system is indicated which uses the above-said algorithm for the algorithm for searching and processing data.

Keywords *Completeness Of Word; Kazakh Language; Intelligent Analysis Of Sentences ; Machine Translation; Information Retrieval.*

1. INTRODUCTION

The process of searching for information represents a sequence of steps leading to a certain result through the system and allowing to evaluate its completeness. Since the user usually does not have comprehensive knowledge of the information content of the resource in which you conduct a search, can assess the adequacy of the query expression, as well as the completeness of the result and can based only on external assessments or on intermediate results and generalizations, comparing them.

The degree of accuracy and completeness of the search depends on how common the terms used in the formulation of the request. It may be wrong to use both the most common terms (the level of information noise increases) and too specific terms (the search completeness decreases). The use of too specific terms may be fraught with the fact that the dictionary of this term may not appear. This process is the same as in machine translation.

The idea of the work may used by any machine translation system. Machine translation as an addition to the classical translation has been the focus of the translation industry. With proper use, machine translation saves time and money. It should be noted that it is suitable for certain types of texts and languages. A good result of translation is provided primarily by two components:

a) well-trained machine translation system
b) subsequent correction of the received machine translation (post-editing).

At post-editing the text received as a result of machine translation is checked and corrected. The purpose of this work is to make only necessary edits to machine translation to improve the quality of the translated text. The time consuming for post-editing depends on the quality of the text received as a result of the machine translation, as well as on the required level of the final translation. Translation of the text for publication requires a greater degree of editing than a translation intended for reading the text [1]. In general, the problem of post-editing machine translation consists of various problems, like morphological, syntactic, spelling, etc. The basis of this idea lies in two stages:

1) determining the wrong or expected wrong words which is incorrect word from the translated text;
2) how to analyze found words.

Since this work is related to post-editing, we need to know clearly what to edit. For this, below, we suggest our approach for finding incorrect words and analyzing them to get the fullness of the meaning of the word related to the context.

At present, many companies like Sanasoft[1], Promt[2], Google[3] are working on the idea of machine translation, and there are also sites with "electronic translators". For example, "*Qazaqstan Respy'blikasy – prezidenttik basqary' nysanyndagy birtutas memleket. Konstity'ci'iag'a sa'ikes, bizdin' el o'zin adam ja'ne onyn' o'miri, ququqtary men bostandyg'y en' qymbat qazynasy sanalatyn demokratialyq, zaiyrly, ququqtyq ja'ne a'leumettik memleket retinde ornyqtyrady.*" The English version of the text "*The Republic of Kazakhstan is a unitary state with the presidential system of government. Under the Constitution, Kazakhstan is a democratic, secular, legal and social state which recognizes the man, his life, rights and freedoms as the supreme values of the country.*" in the Kazakh language was translated by Google machine translation:"*Qazaqstan Respy'blikasy - prezidenttik basqary' ju'iesi bar birtutas memleket. Konstity'ci'iag'a sa'ikes, Qazaqstan adam, onyn' o'miri, ququqtary men bostandyg'y eldin' en' jog'ary qundylyqtary retinde moiyndaityn demokratialyq, zaiyrly, ququqtyq ja'ne a'leumettik memleket bolyp tabylady.*" Even though the basic idea of the translation is preserved, the exact translation still needs quality and subsequent correction(subcrated are different). Therefore, the issue of post-correction of texts on many domestic and worldwide texts is widely considered.

Many scientists are considering different approaches for machine post-editing. Each of them solves a certain problem. For example, Lagard et al. (2009) Based on Roza et al., SMT used static information to post-edit MA output from finished models. Marecek et al. (2011) applied the approach to APE based on the morphological era. Denkovsky (2015) developed a method to integrate the post-edit time of the MA into removing grammar from each input phrase. Recent research has shown that MT plus PE improves the quality of human translation (Fiedererand OBrien, 2009; Koehn, 2009; In addition, DePalma and Kelly (2009), and Zampieri and Vela, also featured production activities in some cases. Recent works have been proposed to use neural networks in many countries. Usually approach consists of two components: the encoder encoding the original sentence and decoding it into the target sentence. Automatic Post-Editing Method Using a New Post-Editing Technique From Statistical Machining Translators offers. The method automatically takes the translation rules as knowledge of translation from the parallel body regardless of linguistic means. Intuitive Translation Guidelines (ICPC) coincides with the global proposal structure and targeted proposal without the need for linguistic means. It also improves translation results by applying translation rules to the translation results obtained through statistical machine translation. Experimental results show the effectiveness of the translation rules for statistical machine translation [2-5]. Their method automatically post-translations result using static data based on machine translation phrases. I would also like to point out that there are a lot of scientific works, but they do not in many respects affect the Turkic language - it is Kazakh and it is complex in processing due to its rich grammatical structure and polysemantic words.

The information system was based on the component software architecture [6]. Component architecture involves the creation of a system consisting of components as reusable nodes. The architecture of information and analytical search engine is presented in Figure 1.1.

[1] http://www.sanasoft.kz/c/ru/node/47

[2] http://www.translate.ru/

[3] https://translate.google.kz/

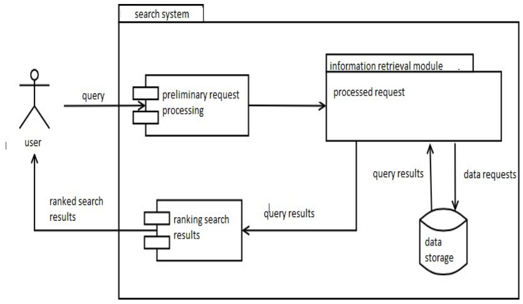


Figure 1.1 - The architecture of information and analytical search engine

The information retrieval system being developed consists of a request preprocessing module, an information retrieval module, a data storage module and a query ranking module. Preliminary processing of a user's request includes checking whether the query matches the search data (the query language, possible errors in the query text, the use of stop words in the query, etc.). The task of preprocessing a request is to make sure that the request has been made correctly and can be processed by the system.

As has been listed there is a lot of work on the search engine and the definition, the use of multi-valued words. But this work is relevant and solves the problems of determining the most appropriate meaning of words when using bulk corpuses and texts. The made system is not ideal and requires even more texts in the Kazakh language and improvement of the program.

The processed request is transmitted to the information retrieval module. The structure of this module is described below. In the process, the information retrieval module accesses the data storage module and receives from it the results containing the answers to the user's request.

The data storage module that is being searched for is a database in which all search data and additional system information necessary for organizing a search are stored (search indexes, an expanded view of search data). The database schema (data storage submodule for search) is presented in Figure 1.2.

Figure 1.2 - Database schema for search

Search data is collected from documents containing search data. Data for the search is stored in text form, in UTF-8 encoding. For each document in the database, an entry is created containing:

- unique identifier of the document;
- document's name;
- the content of the document;
- extended document content (morphologically normalized document content with additional markup).

The fields "content of the document" and "extended content of the document" creates a full-text index, which contributes to increasing the search speed. This index must be rebuilt as data is added or modified. Data storage can be used by any relational database management system (for example, SQLite, MySQL, PostgreSQL, etc.).

The results of requests to the data storage module at the last stage of work are transmitted to the ranking module. Ranking of search results is intended to facilitate the user to work with search results. Currently, search results are ranked in the order in which they are written to the database. At the next stages of the work, it is planned to use ranking algorithms based on the relevance of the search results to the user's initial request.

In order to solve these problems technologically, a flexible information system architecture will be developed. All software modules of the system are interconnected by integration modules (intermediate data storages), which act as connecting links, allowing to obtain a weakly-connected architecture. This design approach allows for relatively easy scalability and upgradeability of modules.

2. FORMAL MODEL OF INFORMATION AND ANALYTICAL SEARCH ENGINE

Initially, in order to understand how the system works to determine the meaning of words and polysemantic words for a search engine for post-development, we’ll take a closer look at the connection with the search engine.

We define that the search result for a system is an undirected graph, where is the set of vertices, and is the set of edges. The vertices are data segments, and the edges of the graph are contextual links between them. Thus, the search engine returns not individual values, but context-related structures from which the user can create an archive of necessary information or select the source of interest of his choice.

We also define the distance function (2.1) on the edge of the graph, which returns a numerical value denoting the level of “similarity” of the vertices and, where, in the search context. A detailed algorithm for calculating the distance is determined by the method of data segmentation and related characteristics of values. In other words, they are individual for each data set and must be defined for each task separately..

$$d_{ij} = \rho(v_i, v_j) \qquad (2.1)$$

Denote the presence of the distance at the edge as $e_{ij} \rightarrow d_{ij}$, where e_{ij} should be read as “edge connecting vertices i and j”. The same applies to the distance.

Note that in order to form a graph $G(V,E)$ using data from heterogeneous sources, reduced to a common standard and stored in the repository S. Define the function of reading data segments (2.2) from a source like:

$$\omega(S,v_i) = \begin{cases} v_j if d_{ij} > 0 \\ 0, else \end{cases} \qquad (2.2)$$

In this case, the time spent on reading, denoted as $\omega(S, v_i) \rightarrow t$.

The goal of the system is to form a graph $G(V,E)$ in the shortest possible time (2.3). Moreover, all vertices must be reachable from any part of the graph. (2.4).

$$\min_t \sum_{i=1}^{n} \omega(S, v_i) \qquad (2.3)$$

$$\exists v_{k_1} v_{k_2} \dots v_{k_s} : (v_{k_i}, v_{k_{i+1}}) \in E, v_k \qquad (2.4)$$

$$\forall v_l, v_p \in V$$

2.1 Development of a model of the information retrieval module of the system

The information retrieval module is a key part of the search engine being developed. The full architecture of the module is presented in Figure 2.3.

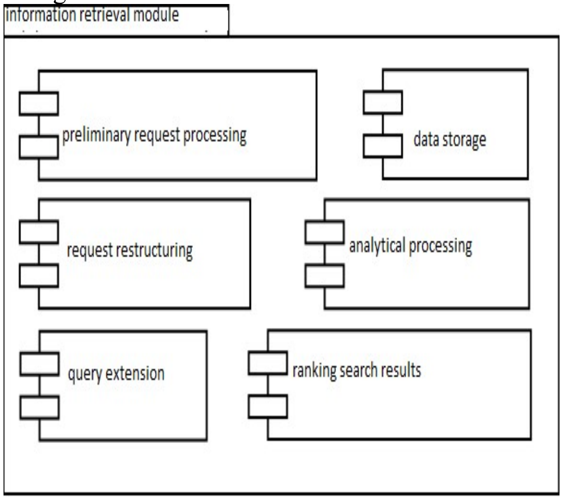


Figure 2.1 - Architecture of the information retrieval module

Information retrieval module of the system include:

- submodule of user request restructuring;
- submodule extension user request;
- submodule of analytical processing of the user's request.

The task of restructuring a user request and expanding a user request is to bring the request into a form that allows for more efficient search of the available data. Restructuring will change the terms from which the request is composed; The extension adds new terms to the query terms provided by the user, which allow to improve the completeness of the search results. Restructuring and expansion of the user's request is one of the

main components of the information retrieval module, which includes:
- morphological normalization;
- the use of a complete system of endings of the Kazakh language, to account for the features of the morphological features of the Kazakh language;
- categorization of queries on the subject of search data.

The morphological normalization is performed by means of the morphological analyzer of the Kazakh language, which is described in section 4. The complete system of the endings of the Kazakh language is used in the form described in [7].

The categorization of queries on the subject of search data is performed using modern classification methods.

2.2 Development of the knowledge base of the information retrieval module of the system

For use in the information retrieval module, knowledge bases of the information retrieval module of the system were developed. Functions of the developed knowledge bases:
- structured data storage, search;
- update and addition of search data;
- building and updating search indexes.

The developed knowledge bases have the following distinctive characteristics:
- consideration of the structure and features of the documents for which the search is carried out;
- morphological normalization of search data to improve the quality of search;
- the possibility of splitting documents into thematic categories;
- accounting semantics of search data.

2.2.1 Development of a knowledge base of idioms for the Kazakh language

When searching, when the input data are idioms, in many cases, search engines issue documents on the result that contain such idioms. Since the phraseological units conveys some value, it would be better for the system to search when issuing not only documents that contain such phraseological units, but also documents that match the value to such phraseological units. To do this, it is necessary to identify and systematize the structure of the type (here you can use the above tagging system), to characterize the specificity of idioms and identify their meaning (using the developed knowledge base of idioms for the Kazakh language).

Phraseological units are very difficult linguistic educations. Their complexity is explained not only by a set of structural types and syntactic models, but both thematic and semantic diversity, an also ability to express the most various emotionally-expressivnye thought shades. It is characteristic that these features allocated phraseological units in many languages, but not in all they are identical.

By idiom can be considered the work of researchers B. Drakshayani and R. Verma. Their work presents approaches for the identification and clustering of phraseological units. For example, in [8], a model of semantic modeling of idioms proposed by the authors is given, documents are clustered based on their meaning using idiom processing methods, semantic weights using a clustering algorithm. The improved quality of creating meaningful clusters was demonstrated and set on three different data sets with documents based on phraseological units using performance indices, entropy and purity. [9] presented an approach to the extraction of phraseological units, which is both domain and language independent, and does not require marking of phraseological units.

There is a lot of phraseological units in Kazakh, and it differs thematic, semantic and structural variety in comparison with other languages. S.'s Kenesbayeva works [10] were defined by B problems of phraseology, criteria of allocation of phraseological units, their classifications were developed. And though separate classification groups of phraseological units were exposed to the scientific analysis, in the Kazakh phraseology is carried so far poorly out the systematized and generalized research in the plan of classification of phraseological units.

In the classification of phraseological units in the existing scientific literature takes into account the sum of all their inherent lexical and grammatical, structural, typological and semantic features. In accordance with them, phraseological units can be assigned to the most diverse classification groups and subgroups: lexical-morphological, structural-typological, syntactic, semantic, subject-specific, stylistic [11].

Considering these groups separately, it should be emphasized that there are no clear boundaries between them, on the contrary, they are closely related, and any grammatical classification of phraseological units cannot but take into account their semantic essence, just like the semantic classification, their grammatical basis. Given these criteria, we have shown below the developed classification of phraseological units.

Classification by structure:

1) in the form of proverbs and sayings. For example, “a’i deytin әzhe zhok, k’oi deytin к’ozha zhok” - this idiom conveys the meaning of “bassyzdyк, ta’rtipscizdik”;

2) in the form of phrases. For example, “barmagin disteledi” - this idiom conveys the meaning of “өkindi”.

Categorization by value:

1) Marking subject. For example, “uki tagylgan кyz” - this idiom conveys the meaning of “aittyrylgan кyz”;

2) Indicating the action of the subject. For example, “zhagasyn u’stau” - this phraseologism conveys the meaning of “tan’dana”;

3) Marking sign of the subject. For example, “U’rip auyzg’a salgandai” - this idiom conveys the meaning “Ademi”.

Phraseological units, regardless of their structural diversity and semantic multiplicity, everywhere appear as relatively solid language units that, in the syntactical plan, act as self-sentences or its members. However, in the system of syntactic constructions, the question of how to introduce phraseological units into one or another syntactic turn is not well understood. In the process of searching for data, if the search query is an idiom, then the following charts will appear.

If the phraseological unit has one value, then the entered query receives this value as a search query. On the basis of the research done, it was found out that in the structure of phraseological units the number of words is from two to six. Next will be presented the developed tagging system for the phraseological units of the Kazakh language, which will be used when searching for the meaning of a given phraseological unit.ph – This symbol indicates that the combination of words is a phraseological unit

<mm> - idiom in the form of a proverb or saying;

<st> - phraseologism in the form of a phrase;

<mn> - a phraseological unit denoting a subject;

<mv> - a phraseological unit denoting the action of the object;

<ma> is a phraseological unit denoting a feature of a subject.

May use the above notation consider a few examples:

Ag’ama zhen’gem say, apama zhezdem say – ph <mm> <ma>;

Aк auyz кyldy - ph <st> <mv>;

Alimynan Aldy - ph <st> <mv>;

Әzhi кyz - ph <st> <mn>;

A’i deytin әzhe , k’oi deytin k’ozha - ph <mm> <mn>. By classification, the above designation is used to determine the structure of the idiom. In further works with an increase in the volume of the idiom base, a tag can be added which describes the length of the idioms (the number of component words). At the moment, from different sources (from books, from Internet resources) have manually collected phraseological units, which are often found.

On the basis of the proposed classification, a knowledge base was developed by the Kazakh idiom, which is replenished with up to 830 records. What allows to improve the quality of information and analytical search engine. The use of phraseological units makes it possible to carry out a semantic search at the level of phrases, in addition to searching at the level of individual words.

3. FINDING THE APPROPRIATE WORDS FOR ANALYSIS

3.1 Combining Algorithm For Definition Of Incorrect Words

As it described in the first chapter, different technologies give different results. For this part, we propose using [12-13] method with own algorithm for determining incorrect words. Our basic idea is aimed at improving the quality of the translated text, and for this we replace the incorrect words we find with a suitable contextual context. In our method, incorrect words are found with the help of any translator from one direction to another and vice versa. The essence of this process, we find incorrect words in parallel for English and Kazakh words as in the face is shown in formula (3.1):

$$S_{e1}\text{----->}\ S_k\text{--->}S_{e2} \qquad (3.1)$$

Where,

S_{e1} --- Source English Sentences

S_k --- Translated Kazakh Sentences

S_{e2}---Target English Sentences

This is a general representation of the model. In this study, the object is the English to Kazakh pair and the parallel corpus for these languages. The peculiarity of the algorithm can be used for any pair to determine the incorrect words.

The main task can be seen in the next.

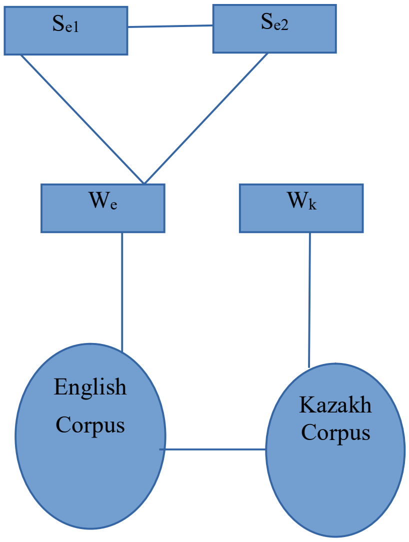


Fig. 3.1.1. Common Form Of Identifying Incorrect Words

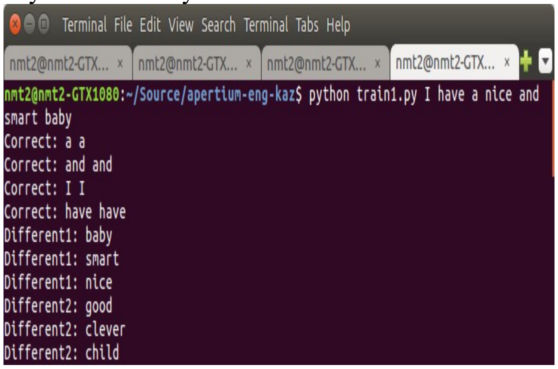
Where,

W_e---English incorrect words

W_k---Kazakh English words

Nowadays, we work with the English text to translate from English into any translator's machine. The main goal is to find the wrong word in any translated Kazakh text, and secondly, to find the directory of the same errors and the third one is to find the correct version of the wrong word. Basically, our work consists of two parts. The purpose of the first part is to identify wrong words.

The purpose of the second part is to analyze all of the wrong words. In the additional material we use the cubic method [14-26] to identify the correct word in the translated sentence and define the correct version of those words for the training set. As for the first question, we work with the results of online machine translation systems as Google. Then we will find the wrong word in the Kazakh language by comparing the original English text and the resulting English text. The Apertium[4] translator system was used for the morphological analysis of the Google translator and sentence to work with the text as above. The basis of the analysis can be done in any translator's system.



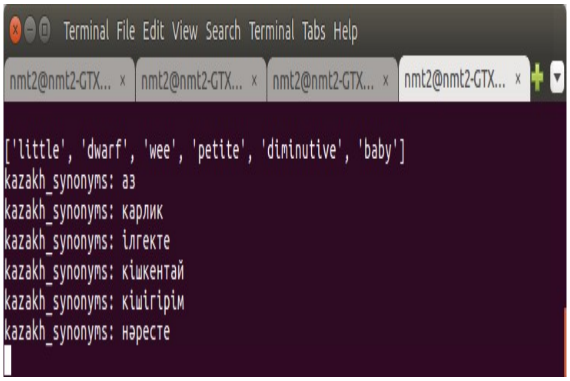
```
nmt2@nmt2-GTX1080:~/Source/apertium-eng-kaz$ python train1.py I have a nice and
smart baby
Correct: a a
Correct: and and
Correct: I I
Correct: have have
Different1: baby
Different1: smart
Different1: nice
Different2: good
Different2: clever
Different2: child
```

Fig. 3.1.2. List Of Found Incorrect Words From The Translated English-Kazakh Sentence

On the second question, we create a dictionary of erroneous words found in translated texts. And analyze them for finding more complicated meaning of the word of the given context. For the analysis, the PyDictionary(Python) library helps us, which provides online synonyms and equivalents of incorrect words. So, the resource used in the English language is used by a Google

[4] http://wiki.apertium.org/wiki/apertium-eng-kaz

translator to translate English words into Kazakh. In this example, one incorrect word with their equivalents received from the online resource is shown in fig.3.1.3:



```
['little', 'dwarf', 'wee', 'petite', 'diminutive', 'baby']
kazakh_synonyms: аз
kazakh_synonyms: карлик
kazakh_synonyms: ілгекте
kazakh_synonyms: кішкентай
kazakh_synonyms: кішігірім
kazakh_synonyms: нәресте
```

Fig. 3.1.3. Different Supposed English And Kazakh Translations Of The Incorrect Word

And thus, we have the meanings of words not only for one language but both in parallel and for the second language. All variants of the necessary words are taken into account.

Wrong word algorithm:

1.$S_{e1} \Rightarrow S_k \Rightarrow S_{e2}$ algorithm is executed.
2.$|S_{e1}|$,$|S_{e2}$ |- is the root of the sentences.
3. $|S_{e1}|\sim|S_{e2}$ |- The roots are mapped to match the unwanted words in temporary memory (excluded as meaning only grammatical parts such as a, the, an,)
4.$|S_k|$- S_k Find the roots of Sk. Algorithm for creating an automatic catalog of alternatives.

Algorithm for creating an automatic catalog of alternatives
2.1 Must be grouped to create an Automated Alternative Catalog. To synchronize, use the online synonyms dictionary (Thesaurus.com-PyDictionary). We memorize all the alternate words in the temporary memory. If we analyze the given example, the two equivalent words found in the online dictionary are:

$|S_{en1}|$ = nice: ['cordial', 'duckly', 'fair', 'friendly', 'good', 'kind', 'lovely', 'superior']
$|S_{en2}|$ = good: ['acceptable', 'excellent', 'exceptional', 'great', 'satisfactory', 'satisfying']

These findings are translated into Kazakh and are equated with (3.1.3) and, if appropriate, will be included in one group of alternatives. If $Se_{n1...nk}$ $Se_{n2...nk}$ & $Sk_{1...n}$ then corresponds, then it belongs to one group. If not, the words will be considered separately.

Parallel corpus for the completeness meaning of the word
After finding these words, we analyze all the found incorrect words with their equivalents for completeness of the meaning of Kazakh words. For example, a simple sentence is taken: *I have a nice and smart baby*. The basic idea is to use any translator shown by the formula(1) is fulfilled and the sentences obtained (two sentences in English) are lemmatized and compared, the incorrect words from these sentences are determined. The lemmatization uses the module of the open machine interpreter Apertium [27] for the English and Kazakh sentence. Another feature of this work after processing the English sentence, we compare the contexts with the Kazakh context in parallel.
After the above algorithm, each word automatically will find its variants from online and additional sources and each will be assigned its own module.

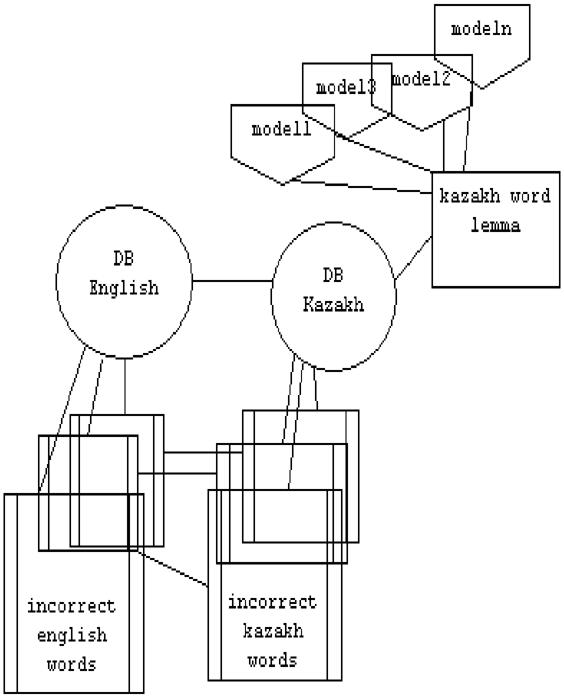


Fig. 3.2. The Basic Model For Determining The Completeness Of The Word For The Kazakh Sentence

In our work we use two English and Kazakh parallel buildings. To find the completeness of a word, the following actions are performed: incorrect words in English are searched from the English database and parallel Kazakh sentences are taken and the second stage takes the equivalents of incorrect words in the Kazakh language and takes the corresponding proposals and as a result all the Kazakh proposals are written in one module. This technique is convenient for any building in the main.

Table 3.2. Comparative Analysis Of Processed Incorrect Words Translated From Different Interpreters For Analyzing Completeness Of The Words.

Source sentences	**Promt**	**Google**
I have a *nice* and smart baby	Mende jaqsy ja'ne aqyldy bala	Mende jaqsy ja'ne aqyldy bala *bar*
I never heard about it before	*Bul turaly buryn* eshqashan *estigen emespin*	*Men* bu'gan *deyin* eshqashan *estimegenmin*
Aknur is a uniqu person	*Aknur - erekshe adamdy ku'raidy.*	*Aqnur - biregei adam*
Students Have started new subject	Studentter jan'a *p'an* bastady	Studentter jan'a *taqyrypty* bastady

The table shows that there are discrepancies in the proposals. Translated sentences with the help of an interpreter Sanasoft require the most extensive analysis of words to improve the quality. Comparatively the Google translator translates the sentences well and does not need any special adjustments. Translator Prompt translates well but still there is a mismatch.

3.2 Practical result

For analysis, we used parallel casings with a volume of 160000 sentences and an online dictionary with a volume of 43,000 and an additional 9,000 Kazakh words of synonyms were included for the completeness of the dictionary.

There are a lot of data in the Kazakh language, but still need to collect data, corps, etc. Dictionaries of different topics in the Kazakh language are enough but the electronic version should be added. But the algorithm works for any pair, but I wanted to indicate that the data was made for the English-Kazakh pair.

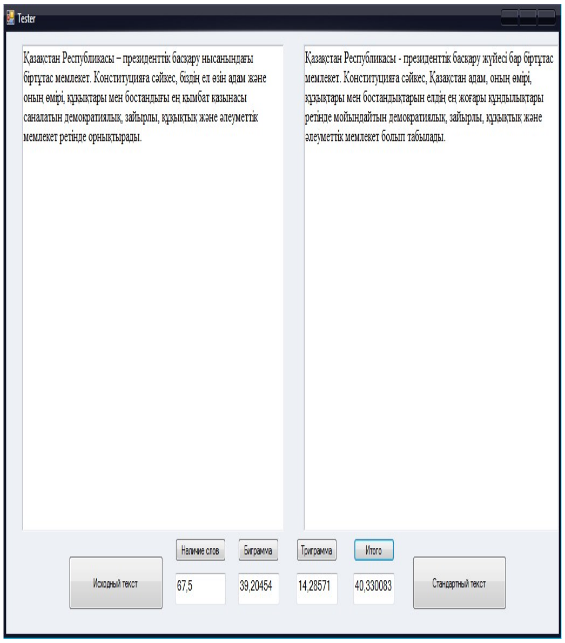
This analysis was conducted on the collected cases in the Kazakh language and multi-valued online and typed by the synonyms themselves. This paper focuses on determining the appropriate meaning of words for polysemantic words as well as to determine the equivalents of the necessary words.

Table 3.3. Percentage Ratio For Using This Methodology For Different Interpreters

Promt	Google	Sanasoft
45%	10%	25%

That is, for the final analysis, 10000 sentences were taken to compare the translated sentence from English to Kazakh language. That is, this proposed system was more useful for all machine translation systems for kaz-eng and eng-kaz and especially for the Sanasoft interpreter. According to the results, it can be seen that Google translates well and the calculation was made according to the statistical method.

Several analyzes were made to verify the quality of the results obtained. Different information on analytics on akorda.kz was received in Kazakh and English languages. The Kazakh text from the site was used as a "golden standard" and subsequent edited text was obtained for comparison purposes.



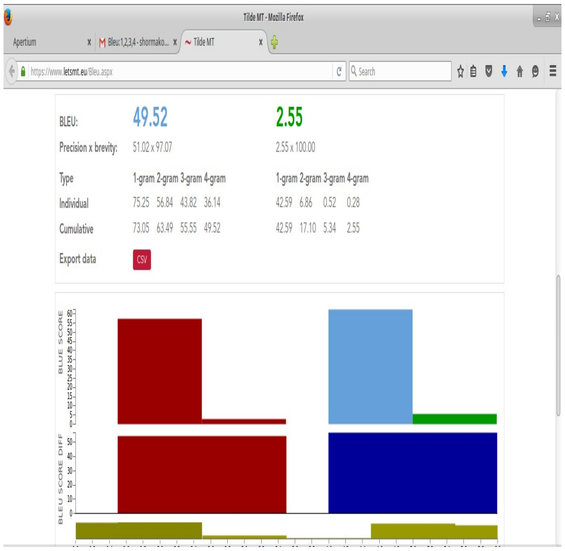
Қазақстан Республикасы – президенттік басқару нысанындағы біртұтас мемлекет. Конституцияға сәйкес, біздің ел өзін адам және оның өмірі, құқықтары мен бостандығы ең қымбат қазынасы саналатын демократиялық, зайырлы, құқықтық және әлеуметтік мемлекет ретінде орнықтырады.

Қазақстан Республикасы - президенттік басқару жүйесі бар біртұтас мемлекет. Конституцияға сәйкес, Қазақстан адам, оның өмірі, құқықтары мен бостандықтарын елдің ең жоғары құндылықтары ретінде мойындайтын демократиялық, зайырлы, құқықтық және әлеуметтік мемлекет болып табылады.

Наличие слов	Биграмма	Триграмма	Итого
67,5	39,20454	14,28571	40,330083

Fig. 3.3. Comparative Analysis Of The Proposal With The Gold Standard Calculated By The Assessment Of Blue Score

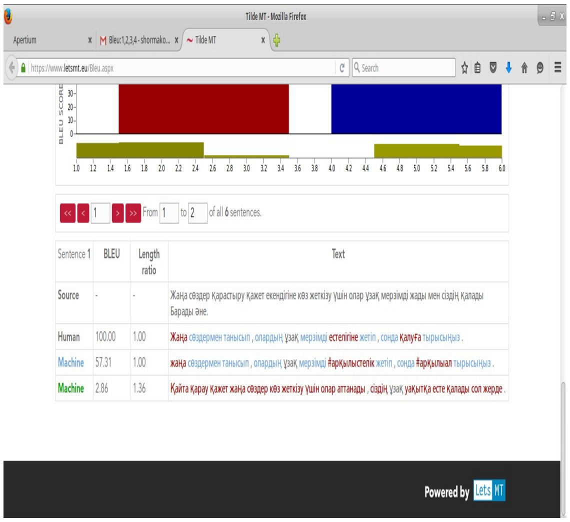
For a clear and clear explanation of the corrected Kazakh texts, another Internet source was used, where the analysis from each sentence is clearly and distinctly visible. 49.52 Bleu calculated for the whole sentence, and figure 2.55 is calculated for bigrams, etc.



BLEU:	49.52	2.55
Precision x brevity:	51.02 x 97.07	2.55 x 100.00
Type	1-gram 2-gram 3-gram 4-gram	1-gram 2-gram 3-gram 4-gram
Individual	75.25 56.84 43.82 36.14	42.59 6.86 0.52 0.28
Cumulative	73.05 63.49 55.55 49.52	42.59 17.10 5.34 2.55
Export data	CSV	

Fig. 3.4. Comparing Bigrams, Etc. For Several Kazakh Sentences

For clarity, you can see in the following figure the differences between sentences marked with different words. Thus, we find the difference between the gold standard and the translated sentence.



Sentence 1	BLEU	Length ratio	Text
Source	-	-	Жаңа сөздер Қарастыру қажет екендігіне көз жеткізу үшін олар ұзақ мерзімді жады мен сіздің қалады Барады әне.
Human	100.00	1.00	Жаңа сөздермен танысып , олардың ұзақ мерзімді естелігіне жетіп , сонда қалуға тырысыңыз .
Machine	57.31	1.00	жаңа сөздермен танысып , олардың ұзақ мерзімді #арқылыстелік жетіп , сонда #арқылыал тырысыңыз .
Machine	2.86	1.36	Қайта қарау қажет жаңа сөздер көз жеткізу үшін олар аттанады , сіздің ұзақ уақытқа есте қалады сол жерде .

4. ACKNOWLEDGMENTS

This work was done with funding **Project IRN** AP05132950 "Development of an information-analytical search system of data in the Kazakh language" Ministry of Education and Science of the Republic of Kazakhstan.

5. CONCLUSION

In conclusion, in this paper, the author presented combining method that automatically determines the incorrect words and definitions of the completeness of words for certain contexts that improve the quality of the translation. In the conclusion of this work translations of the sentences from English into Kazakh were given. In the proposed system, Kazakh words were trained to improve the quality and blot out words in context. Analyzing inconsistent, incorrect, questionable words of several translators, comparative results were obtained in percentage. As well as the work is associated with a search engine as when searching requires additional parameters and estimates to select the best indicator. The meanings and meanings of words, especially for the Kazakh sentences, are distinctive and have their own specificity; therefore, these studies improve the quality of finding and completeness of the meaning of words for a given language.

The translators of the English-Kazakh couple are developing better and faster every day. But nevertheless, it is still necessary to work on semantic, syntactic sections for the best result.

One of the most important elements influencing the results of information retrieval is the thesaurus of keywords, which includes the expansion of the subject area due to synonymy and the formation of the thesaurus of synonymy on this basis. In the 1970s, thesauri began to be actively used for information retrieval tasks. In such thesauri words are mapped to descriptors through which semantic links are established. The first modern English thesaurus was created by Peter Mark Roger in 1805. It was published in 1852 and has since been used without reprints. There are also electronic dictionaries of synonyms and thesauri of the English language, etc. [28-45]. Unfortunately, for Turkic languages, electronic linguistic resources are not publicly available, which would allow to apply artificial intelligence in various applied tasks. Therefore, it is not a little important to use the knowledge base of synonymy to search for and complete the meanings of words in the Kazakh language.

The nature of the translator and the requirements for it have changed dramatically over the last 10-15 years. First of all, translation of scientific and technical, official and business documents has changed. It is not enough to translate text using the computer as a writer today. The Client expects the translator to comply with the original nature of the original document and satisfy the country's standards. The interpreter also has to be able to effectively apply orders made on that subject, and the employer is also encouraged to save time and money in the translation of duplicate text or fragments of text. The strict and often contradictory interpretation of the terms implies that the translator is not only mastering native languages and other foreign languages and the subject area chosen, but also with the use of modern computer technologies.

REFERENCES

[1] Giselle de Almeida, Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages. Thesis, School of Applied Language and Intercultural Studies Dublin City University,p 8–48 (2013). DOI= http://doi.acm.org/10.1145/161468.16147.

[2] Philipp Koehn. A process study of computer-aided translation. Machine Translation, 23:241–263, 2009a. ISSN 0922-6567. doi: 10.1007/ s10590-010-9076-3. URL http://dx.doi.org/10.1007/s10590-010-9076-3.

[3] Rudolf Rosa, Ondˇrej Duˇsek, David Mareˇcek, and Martin Popel. Using parallel features in parsing of machine-translated sentences for correction of grammatical errors. In Proceedings of Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6), ACL, pages 39–48, Jeju, Korea, 2012a. ACL, ACL. ISBN 978-1-937284-38-1.

[4] Michael Denkowski, Alon Lavie, Isabel Lacruz, and Chris Dyer. 2014. Real time adaptive machine translation for post-editing with cdec and transcenter. In Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation, pages 72–77, Gothenburg, Sweden, April. Association for Computational Linguistics.

[5] Aziz, Wilker, Maarit Koponen and Lucia Specia (2014). "Sub-sentence Level Analysis of Machine Translation Post-editing Effort." O'Brien et al. (2014), 170–199.

[6] ISO 25964-2:2013 Information and documentation – Thesauri and interoperability with other vocabularies. Interoperability with other vocabularies. – 2013. - Part 2. – 150 p.

[7] Tukeev UA, Turgynova A. Morphological analysis of the Kazakh language on the basis of a complete system of endings // Proceedings of the Intern. conf. in computer and cognitive linguistics (TEL-2016). - Kazan, Tatarstan, 2016. - p. 225

[8] Wang, Jing, and Yuchun Guo. Scrapy-based crawling and user-behavior characteristics

analysis on taobao // 2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. IEEE. – 2012. – P. 44-52.

[9] Myers, Daniel, and James W. McGuffee. Choosing scrapy // Journal of Computing Sciences in Colleges. – 2015. 31, No. 1. – P. 83-89.

[10] B. Drakshayani, E. V. Prasad. Semantic Based Model for Text Document Clustering with Idioms // International Journal of Data Engineering (IJDE). – 2013. Vol.4, Iss.1. – P.1 -13.

[11] R. Verma, V. Vuppuluri. A New Approach for Idiom Identification Using Meanings and the Web // Proceedings of Recent Advances in Natural Language Processing. – Hissar, Bulgaria, 2015. – P. 681-687.

[12] Kenesbayev SK, Phraseological dictionary of the Kazakh language. - Alma-Ata: Science, 1977. - p. 711.

[13] Vinogradov V.V. On the main types of phraseological units in the Russian language // Selected Works. Lexicology and lexicography. - M., 1977. - 135 P.

[14] Miquel Esplà-Gomis, Felipe Sánchez-Martínez, Mikel L Forcada, "Using machine translation in computer-aided translation to suggest the target-side words to change", in *Proceedings of the 13th Machine Translation Summit* (September 19-23, 2011, Xiamen, China) , p. 172-179

[15] Espla-Gomis, M., Sanchez-Martinez, F., & Forcada, M. L. (2015). Using on-line available sources of bilingual information for word-level machine translation quality estimation. In Proceedings of the 18th Annual Conference of the European Association for Machine Translation, pp. 19{26, Antalya, Turkey.

[16] Assem Abeustanova,Ualsher Tukeyev, Automatic Post-editing of Kazakh Sentences Machine Translated from English, Studies in Computational Intelligence 710 Advanced Topics in Intelligent Information and Database Systems, 283-295pp, Springer 2017

[17] Python Homepage, https://pypi.python.org/pypi/ PyDictionary/1.3.4, last accessed 2018/04/25

[18] Apertium Homepage, http://wiki.apertium.org/wiki/ Main_Page, last accessed 2018/04/25

[19] Espla, M., Sanchez-Martinez, F., & Forcada, M. L. (2011). Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited. In Proceedings of the 15th Annual Conference of the European Association for Machine Translation, pp. 81{89, Leuven, Belgium.

[20] https://en.wikipedia.org/wiki/Principle_of_ maximum_entropy

[21] Translator Promt, http://www.promt.ru

[22] Translator Google , https://translate.google.kz/#kk/en

[23] http://www.krugosvet.ru/enc/gumanitarnye_ nauki/lingvistika/MASHINNI_PEREVOD.html

[24] Koehn, Statistical Machine Translation, p.113 , http://www.statmt.org/book/

[25] http://bazhenov.me/blog/2013/04/23/maximum-entropy-classifier.html

[26] Translator Yandex , https://translate.yandex.kz/

[27] http://wiki.apertium.org/wiki/Main_Page

[28] "Senez, Dorothy. "Post-editing service for machine translation users at the European Commission". Translating and the Computer 20. Proceedings from Aslib conference, 12 & 13 November 1998; Vasconcellos, M. and M. Léon (1985). "SPANAM and ENGSPA: Machine Translation at the Pan American Health Organization". Computational Linguistics 11, 122-136".

[29] "Allen, Jeffrey. "Post-editing", in Harold Somers (ed.) (2003). Computers and Translation. A translator's guide. Benjamins: Amsterdam/Philadelphia, p. 312".

[30] "Loffler-Laurian, Anne marie. (1986). Post-édition rapide et post-édition conventionelle: deux modalities d'una activité spécifique¨. Multilingua 5, 81-88".

[31] "Wagner, Elisabeth. "Rapid post-editing of Systran" Translating and the Computer 5. Proceedings from Aslib conference, 10 & 11 November 1983".

[32] Green, Spence, Jeffrey Heer, and Christopher D. Manning (2013). "The

Efficacy of Human Post-Editing for Language Translation". ACM Human Factors in Computing Systems.

[33] Plitt, Mirko and Francois Masselot (2010). "A Productivity Test of Statistical Machine Translation Post-Editing in A Typical Localisation Context". Prague Bulletin of Mathematical Linguistics. **93**: 7–16.

[34] Marcello Federico, Alessandro Cattelan, and Marco Trombetti (2012). "Measuring user productivity in machine translation enhanced computer assisted translation". Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA), San Diego, CA, October 28 – November 1.

[35] Läubli, Samuel, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk (2013). "Assessing post-editing efficiency in a realistic translation environment" (PDF). Proceedings of the 2nd Workshop on Post-editing Technology and Practice. pp. 83–91.

[36] "Translation Productivity Tools for Businesses". Pairaphrase. Pairaphrase Blog. Retrieved 3 January 2018.

[37] *Hutchins, John (1995).* "Reflections on the History and present state of machine translation"

[38] "Unbabel Launches A Human-Edited Machine Translation Service To Help Businesses Go Global, Localize Customer Support"

[39] Information technology information retrieval //http://inftis.narod.ru/is/is-n8.htm/: 10.09.2018.

[40] General principles of organizing information retrieval on the Internet //https://sites.google.com /site/gisciencepsu/in-the-news/ : 10.09.2018.

[41] GOST 7.73-96. System of standards on information, librarianship and publishing. Search and distribution of information. Terms and Definitions. -1998. -13 P.

[42] Abiteboul S., Buneman P., Suciu D. Data on the Web: From Relations to Semistructured Data and XML.2014.-260 p. // *homepages.dcc.ufmg.br/~laender/material/ Data-on-the-Web-Skeleton.pdf /:* 10.05.2018.

[43] Buneman P., Davidson S., Fernandez M., Suciu D. Adding Structure to Unstructured Data // In Proceedings of the International Conference on Database Theory. - 1997. -P. 234-240.

[44] Sint R., Stroka S., Schaffert S., Ferstl R. Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis // Proceedings of the Forth Semantic Wiki Workshop, Heraklion. - Crete, Greece, 2009. -P. 56-60.

[45] Masterman M. Semantic message detection for machine translation, using an interlingua // Proc. International Conf. on Machine Translation, 1961.– P. 438-475