Automatic Post-editing of Kazakh Sentences Machine Translated from English

Assem Abeustanova and Ualsher Tukeyev

Abstract Automatic post-editing of sentence from one language to another language is closely connected with a machine translation. Machine translation is a modern tool that helps to speed up the translation process and to reduce its cost. But still it is a problem to get a correct translation. Therefore now it is active investigates automatic post-editing improved quality of translation. In this paper is proposed method of automatic post-editing based on two stage procedure: on the first is defined incorrect words in a text by using translation memory (TM) technology, on the second stage is defined what alternative translation word is more better for defined incorrect words on first stage by using of maximum entropy model.

Keywords Automatic post-editing · Translation memory technology · Maximum entropy model

1 Introduction

As for machine translation, it should be noted that the computer does not have mind. It does not understand the nuances of language, hints in the text. Each new construction, phrase, idiomatic expression should be provided by the programmer and are included in the program. Depending on the style and purpose of the text, the same word can often have several meanings. In our time, machine translation software can recognize the style of text, and, depending on this, select the necessary dictionary. Also, the program itself can offer several translations to translator. Therefore, considering all these nuances the work was done.

The basic idea of this paper is to find the incorrect words in the Kazakh sentences that translated from English sentences and automatic correct the incorrect

A. Abeustanova (✉) · U. Tukeyev
Al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail: shormakovaassem@gmail.com

U. Tukeyev
e-mail: ualsher.tukeyev@gmail.com

D. Król et al. (eds.), *Advanced Topics in Intelligent Information and Database Systems*, Studies in Computational Intelligence 710,
DOI 10.1007/978-3-319-56660-3_25

Kazakh words. As a result, the resulting work we get corrected right sentence in Kazakh sentence. For the finding in translated text a incorrect words is proposed to use translation memory technology [1] and for automated correct of incorrect words is proposed to use maximum entropy model [2]. The system was previously "trained" on the parallel corpus and then applied to "working" texts. Due to the preliminary "training" it is possible to achieve higher accuracy of the translation, more suitable terminology and to reduce the costs of post-editing.

2 Related Works

Among the many online translators PROMT [3] and Google [4] should be distinguished. These programs are used for comparative analysis. In order to understand the principles of actions and dictionary usage quality and grammar analysis, also the quality of the translation, the text was translated. Translation was carried out in the English-Kazakh direction (a text consisting of 51 words). The following two criteria were used in the analysis: (1) the proper selection of words's meaning by the system (lexical level), (2) the accuracy of matching words in a sentence (grammar level, the coordination of words in a sentence in gender, number, face, case, and punctuation). Take the following passage of a text as an example: "It would be hard to imagine a more evil piece of work than Robert Alton Harris. After a lifetime of vicious, random crime, in 1979 in California he murdered two teenage boys in cold blood for their car. As he drove away, he finished off the cheeseburgers they had been eating". Translation made by automatic translation system PROMT: "Edi elestetu kiyn astam zloe tuyndysy karaganda,Robert Al'ton Harris. Uzak jyldardan kein katygez kylmys, 1979 jyly Kaliforniada ol oltirgen eki jasospirim suyk kan oz avtokoligi ushin. Ol ketip kaldy, ol dobil chizburgery olar shyrsha.". The system has successfully coped with the search for the English equivalent of "in cold blood". Pay attention to the "workpiece". It is easy to understand that this error is caused by polysemy of words "piece" and "work". Furthermore, it was absolutely incorrect to translate the phrase "oz koligi ushin".

Moreover the author of the text within the meaning of the first sentence refers to the detail, the man would never have confused whereas it is animate or inanimate. The phrase "finished off the cheeseburgers" was translated by PROMT as "dobil chizburgery". But this phrase is unacceptable in the Kazakh language. A significant drawback is that in the output language word order is almost always the same as the input. Translation made by Google's online translator: Ol Robert Harris Alton Karaganda jumys negurlym jaman boligin elestetu kiyn bolar edi. Yzaly, kezdeisok kylmys omir kein, 1979 jyly Kaliforniada ol olardyn avtomobil'ge arnalgan suyk kan eki jasospirim ul oltirgen.Ol ketip, ol olar jep boldy chizburgery oshiru aiaktaldy." in contrast to the PROMT Google has translated idiom "in cold blood" and issued for translation "suyk kan", i.e. used the literal translation. Google also repeated the error in the PROMT translation "for" as "ushin". Instead of using the previous program's dobil chisburgery" Google used "jep boldy chizburgery", that is

better. In the English-Kazakh translation online program PROMT made 6 errors (2 lexical and 4 grammatical). And Google also made 6 errors (3 lexical and 3 grammatical). From this study, it follows that the least number of errors in the translation from English into Kazakh was made by PROMT system. After completing the analysis of above mentioned modern systems of machine translation, we noticed that every machine translation system has its strengths and weaknesses. At this stage, the machine translation systems can not exist without a man, because we noticed that the translation is not perfect. Therefore, if we want to get high-quality translation, post-editing by a man is necessary [5].

The work of the Spanish group [6] focused on the sub-segments of the qualitative evaluation of machine translation (MTQE) at the word level. The main advantage of the verbal level MTQE is that it allows not only to estimate the effort needed for post-editing output from the MT system, and which words need to be edited after as the guidance for post-editors. In this article there is used the black box of bilingual resources from the Internet to the level of MTQE words. Namely, they combine two online MT systems, Apertium and Google Translate, and bilingual concordancer Reverso Context3 to detect a sub-segment of correspondence between sentences S in the original language (SL) and the translation of T hypothesis in the target language (TL). For this purpose, both S and T is segmented into overlapping sub-segments of variable length and they are translated to TL and SL, respectively, using a bilingual sources mentioned above. These matching sub-segments are used to retrieve a collection of functions that are then use a binary classifier to determine the word to be edited later. Their experiments confirmed that their method gives results comparable to level of technique, using significantly less capacity. Moreover, considering the fact that there are used (online) resources in their method, which are publicly available on the Internet and as soon as the binary classifier will be trained it can be used at the level of words on the MTQE doing new translations. Work inspired by the work Esplà-Gomis et al. [7], in which several MT-line systems are used to assess the quality of speech at a translation memory (TM) based on the automated translation system tasks. In the the paper of Esplà-Gomes and others (2011), taking into account the European Parliament (S, T) suggested the translator to the SL segment S, MT is used for the translation of the S sub-segments in the TL and TL sub-segments of the T sub-segment at the Sea Level is obtained through MT, which occur both in S and T are a evidence to the fact that they belong to. Alignment between S and S, together with a sub-segment transfers between S and T help to determine which words in T should be modified to obtain T, the desired translation S. There is considered the use of a general-purpose machine translation (MT) to assist users of computer translation (the CAT) based on translation memory systems (TM) to determine the target word in the target sentences to be changed (or changed or deleted) or kept unaltered the problem referred to as guidelines for the maintenance of the word. MT is used as a black box to align source and target sub-segments in the translation of units offered to the user. The source of language (SL) and the target language (TL) segments in matching TUs are segmented in overlapping sub-segments of variable length and translated into machine TL and SL, respectively. Bilingual subsegments are obtained and agreed

between the SL segment in the segments TU and parts of translation are used to create functions, which are then used by binary Classifier to determine the target word to be changed, and those that will be saved unedited. Two approaches are presented in this paper: one using words to support recommendation system, those which can be trained on TM are used with CAT and more basic approach, which does not need any preparation. Experiments are hold by modeling text translation to several languages from corpuses, belonging to different aspects and with usage of 3 systems of MT. In this paper, by comparing previous works we suggest use and compose alignment technology with the method maximum entropy for automated post-editing of sentence from English to Kazakh to improve the quality of translation.

3 Method

3.1 Combining of Translation Memory Technology and Maximum Entropy Method

The approach we proposed consists of two modules. The function of defining incorrect words and post-editing of these incorrect words make analyses of words for post-editing and the module of text for the usage of carried out analysis. The method maximum entropy used for automatic post-editing of words and phrases of English-Kazakh sentences. The main objective of the training model Stage 1 is the definition of incorrect words and need to adjust these words. Initially we take the English sentences and translate them through a translator Apertium and in addition to manually write the expected correct translation of the English sentences. Eventually we have two files: the Kazakh sentences from translator and translated by person. These sentences of two files (translation of Apertium and the expected correct translation) are aligned by matrix phrase alignment method (Koehn, Statistical Machine Translation, p. 113) [8], and inappropriate segments and words of translated Kazakh text are determined what defined.

The main objective of Stage 2 is adjustment of sentence with predefined incorrect words. There are used a small number of parallel sentences on two languages (about 100,000) for initial data analysis. These sentences are used for building tables for training the system. There is we use method maximum entropy for generating cube tables. That is, the system is trained by cube tables created in advance. The table is constructed based on a bilingual parallel corpus and bilingual dictionary.

Indicators which give one unit as the remainder of the division by the number of classes, triggered (return true) only in the first grade, multiple to the two for the second, etc. This approach is not mandatory for the implementation of the classifier, but in order to understand the theory it is important to understand the difference between a sign and indicator, as well as differences in their numbering.

Classification takes place by the formula:

$$p(c \mid d, \lambda) = \frac{\exp \sum_i^{n \times k} \lambda_i f_i(c, d)}{\sum_{\tilde{c} \in C} \exp \sum_i^{n \times k} \lambda_i f_i(\tilde{c}, d)} \quad (1)$$

In this formula [9]:

- f_i—i-th classification indicator (0 or 1);
- λ_i—the weight of the i-th classification indicator f_i;
- c—class hypothesis;
- C—the set of all possible classes;
- D—classified document.

Each indicator has a weight of f_i λ_i, which describes the relationship between the relevant classification criterion and class. The greater the weight, the stronger the connection. Thus, the numerator describes the exponential weights for class-hypothesis, and the denominator normalizes the value of the unit. The most difficult part of this formula—a set of weights λ.

3.2 Realization of Maximum Entropy Method

Maximum entropy method is very useful as well as the generation of tables and to determine the most probable word. That is, when using this method is given below the above description and in the end we get the equivalent word with the closest meaning to the context. The result of this attitude is not just a classification decision, it is the probability for a given class. One of the advantages of this classification is that it is much more accurately models the probable distribution of the classes. Using a machine translation from English into Kazakh translated text should be edited if there incorrect word in a sentence.

Then after finding the incorrect words in a sentence, alternative look at their translation so that we can insert and give post-edited correct translation and further is used on the basis of the method of maximum entropy. General description of the method is as follows:

$$f_i^j = \begin{cases} 1, & if\ d = w_i, c = AW_j \\ 0, & in\ other\ cases \end{cases} \quad (2)$$

where AW_j—alternative word (class), d—classified word.

In order to correct incorrect words multivalued dictionary databases and TM (Translation Memory) are used. The following items are shown as an example (Table 1).

Small case of sentence in 1235 was taken for accurate analysis. From this body it was defined that the word “ana” is incorrect and its equivalents were found:

Table 1 Two sense of the same words

Alternative words	Collocations
ana	anama kyzykty kitap al, ana jaksy koredi
mama	mamasynyn kuanyshyna ainalady
ene	enesi pisirgen
sheshe	sheshem balish pen bir kuty maimen
apa	apasy balish pisirip

1 *Anasy* ony ote jaksy koredi eken. (The *mother* was very fond of it)
2 Bir kuni *apasy* balish pisirip, kyzyna keshikpeuin aitady. (One day, her *mother* bake a cake, the daughter of late say)
3 Ogan *enesi* pisirgen balishinen jane bir kuty mai alyp bara jatyrmyn. (It'm going to my *mother's* bälişinen and a bottle of cooking oil)
4 Sen myna jolmen bar, men *ana* jolmen jurein. (You are this way, and the mother *in* a way)
5 Men sizge *sheshem* balish pen bir kuty maimen jiberdi. (I sent you my cake and a bottle of oil)
6 Uly akesinin maktanyshyna, *mamasynyn* kuanyshyna ainalady. (The pride of his father's, *mother's* joy becomes)
7 Sen *anama* kyzykty kitapty al. (You have an interesting book and a *mother*)

We used maximum entropy:

$$f^1 = \begin{cases} 1, & if\ d = ``f_3 \wedge f_4 \wedge f_6", c = AW_1 \\ 0, & in\ other\ cases \end{cases} \tag{3}$$

$$f^2 = \begin{cases} 1, & if\ d = ``f_1 \wedge f_2 \wedge f_5", c = AW_2 \\ 0, & in\ other\ cases \end{cases} \tag{4}$$

$$f^3 = \begin{cases} 1, & if\ d = ``f_2 \wedge f_5", c = AW_3 \\ 0, & in\ other\ cases \end{cases} \tag{5}$$

$$f^4 = \begin{cases} 1, & if\ d = ``f_2 \wedge f_3 \wedge f_4", c = AW_4 \\ 0, & in\ other\ cases \end{cases} \tag{6}$$

$$f^5 = \begin{cases} 1, & if\ d = ``f_4 \wedge f_5", c = AW_5 \\ 0, & in\ other\ cases \end{cases} \tag{7}$$

According to these rules, Table 2 was built.

Table 2 Calculations of probability for each case

		f_1	f_2	f_3	f_4	f_5	f_6
AW_1	f^1	0	0	1	1	0	1
AW_1	Weight	–	–	1/7 = 0.142	2/7 = 0.285	–	1/7 = 0.142
AW_2	f^2	1	1	0	0	1	0
AW_2	Weight	5/7 = 0.714	1/7 = 0.142	–	–	1/7 = 0.142	–
AW_3	f^3	0	1	0	0	1	0
AW_3	Weight	–	3/7 = 0.428	–	–	2/7 = 0.285	–
AW_4	f^4	0	1	1	1	0	0
AW_4	Weight	–	4/7 = 0.571	1/7 = 0.142	1/7 = 0.142	–	–
AW_5	f^5	0	0	0	1	1	0
AW_5	Weight	–	–	–	3/7 = 0.428	5/7 = 0.714	

As a result of calculating, the probability of incorrect words separated by parts of speech, were as follows:

$$P(AW_1) = 0.427$$
$$P(AW_2) = 0.998$$
$$P(AW_3) = 0.713$$
$$P(AW_4) = 0.855$$
$$P(AW_5) = 1.142$$

That is, the maximum entropy method selects value 1.142, and it gave better full decision using of the features and weights of different alternative words. That is only particular parts of sentence and the maximum meaning of probabilities, whereas after used the complemented generating cube tables we proposed considers the contexts of used incorrect words. That is, taking into account not only the parts of speech and probability of necessary words, but also the meanings of each needed word in the text. And this method is used in the Functions of defining incorrect words and post-editing of these incorrect words in Stage 2 in the construction of tables for each incorrect word.

4 Experimental Results and Discussion

4.1 Description of Function of Defining Incorrect Words and Post-editing of Them

The main work consists of two stage (modules). First stage is to find the right word in the Kazakh language translated from a person entered any English sentence. This part of the find and mark the incorrect words or segments of the sentence in the Kazakh language is based on the method of translation memory. The Function of defining some incorrect words and post-editing of these incorrect words Stage 1—is associated with a list of incorrect words, that is, if to make a detailed description, it is associated with the File consisting of three types of sentence (English sentences, Kazakh sentences, correct Kazakh sentences) obtained after algorithms of post-operators. In each line it looks as follows: *interesting subject*, *kyzykty sabak*, *kyzykty pan* etc. We find the wrong words, divide them, find the roots and compare them, that is, if to take from the example it will be *sabak*. Post-operator algorithm was edited and as a result defines only one wrong word and is written with a list of polysemantic words in the file. So, as a result, when there will be added new sentences in the initial three files, it automatically appears in this new file. And since we have to cover a list of incorrect words as more as possible we consider polysemantic words too. Morphological Analyzer Apertium is intended to Stemming algorithm for the Kazakh language to divide words from the roots and ends.

Function of defining incorrect words and post-editing of the incorrect words Stage 1—After determining the incorrect words we work with tables. That is, this table is ready for any new sentence. The wrong word found in defining incorrect words. We do following to reach readiness:

Stage 1:

1. search this wrong Kazakh word from the English-Kazakh dictionary, remember all the translations of this word and english version respectively.
2. then look for the english basis of the sentence (25,000 at the moment, it may be more) all sentences with this English word and consider Kazakh translation of these sentences.
3. make a table for each wrong word, and name each table by the English version of the word. That is, if the word *sabak* (subject) is incorrect and it occurs as a subject in the dictionary, then the table will be named *subject* for this word. And these all tables were saved as a file for each incorrect word. We remember all the synonyms of words found in English-Kazakh dictionary and write words nearby (only the roots of the word) from the found sentences which relate only to the wrong word, morphological analyzer of Apertium is used for these purposes. And calculate for each case how many times they appear in the 100,000 base of sentences. You can add even more sentences for corpus. And so it turns out a lot of tables with ready frequencies and incorrect words.

4.2 Description of Phase TEST

On the second stage, we fix already found in the first paragraph of incorrect words and for this purpose the generation of cube tables based on the maximum entropy method. That is found objectionable words are corrected based on the cube tables that have previously been generated. To there exists English-Kazakh dictionary which is required to determine the equivalents of the English translation of these words or segments that have been found incorrect. And they found the English and Kazakh equivalents of these words are searched for from a bilingual corpus. And this most communication takes place with the context and it can determine the meaning of a proposal based on context and word. From parallel English-Kazakh corpus considering Kazakh sentences computed by table of the cube is below described in more detail:

At this stage, we use the previous module to translate any incoming sentences in English.

$$\text{Sentence}_{\text{english}} \rightarrow \text{Sentence}_{\text{Kazakh}} \rightarrow \text{Sentence}_{\text{correct}}$$

1. Yandex [10] translator translates automatically.
2. The function phase of determining incorrect words and post-editing of these incorrect words, that is, the work of Stemming Algorithm is performed, and search for incorrect words from incorrect bad_slova.txt file from already finished table to calculate the probability for each found incorrect word.
3. probability algorithms are used.

When all the words are found from the table it is necessary to take into account the translated sentence that we correct. After that is we are not going to use an entire file, which means that we will take those words that are used only in this sentence. For example: the translated text will be: *Sabak kesh boldy*.

(The lesson was late.) That is we found incorrect word *sabak*. (Subject has some synonyms as lesson, object etc.) Now if you look at the table, let's say the file looks as follows: (The data are taken from one table)

3: men; sabak
2: men; takyryp
1: men; sub'ekt
1: keshe; sabak
4: keshe; takyryp
1: keshe; sub'ekt
1: ol; sabak
1: ol; takyryp
1: ol; sub'ekt
1: bol; sabak
1: bol; takyryp
1: bol; sub'ekt

Only those words which are found in translated sentence are used from this file. *Yesterday was a subject*, Apertium [11] translated as: ***Keshe sabak boldy*** and so we take only words *keshe* and *bol* with polysemantic words and calculate the probability for these words only:

1: keshe; sabak
4: keshe; takyryp
1: keshe; sub'ekt
1: bol; sabak
1: bol; takyryp
1: bol; sub'ekt

It is Calculated by the formula: $P(s) = P(s_1) + P(s_2) + \cdots + P(s_n)$ and use the probability to above words with their sentences, *Subject* has some synonyms as *lesson*(*саба*), *object*(*ма ырып*), *subject* (*субъект*):

$$P(\text{sabak}) = 1/25000 * 1/25000 = 0.00004 * 0.0004 = 0.0000000016$$
$$P(\text{takyryp}) = 4/25000 * 1/25000 = 0.00016 * 0.00004 = 0.0000000064$$
$$P(\text{sub}'\text{ekt}) = 1/25000 * 1/25000 = 0.0000000016$$

And select the maximum value and get a second probability which is *takyryp*, and paste this value into the sentence.

4. In order to paste the calculated right word it is necessary to find in which endpaste the word in sentence and the Morphological Analyzer Apertium is used. And as a result we get the full correct sentence with post-edited word or words.

The above mentioned and described an example of the result of the program is aimed at, that is, the phase of testing. That is, as a result we get the right corrected Kazakh sentence, and this sentence from English into Kazakh language translated through any interpreter translated sentence on working consistently with Stage 1 and Stage 2 to obtain high-quality translation.

In the development of the algorithm it is required to supplement the data, that is, the more sentences the more accurate information about post-editing of sentences. Base TM would be better updated and added.

Note, this may be:

1. if at least one root of the word is not there in bad_slova.txt we don't look to this sentence and explain that this word has no incorrect words.
2. it is possible to find this word in bad_slova.txt but not found in the list of tables from the stage of defining incorrect words and post-editing of them, we should take into account these words and algorithm of building the tables should add these new words.
3. it is possible that the sentence have several incorrect words, then we take all the incorrect words.

As a result of the made work using this technique it has been made the small analysis from small the sentence (the 100th offer). Different translators for check of the first stage have been used where is defined incorrect words or segments. By results of the carried-out small analysis it has been shown the following results.

As the result shows using the first stage of the offered method above described from these three translators the modified number of words corresponding to tables have been found. According to Table 3 it is visible that it is possible to find words from a context of which it is necessary to correct using offered by us by method translation memory. But as this inexact assessment, we can't be reliable that

Table 3 Percentage indicator finding of the translations of incorrect words from several translators

Google	Apertium	Promt
12%	10%	15%

Table 4 Percentage indicator coincidence of incorrect words between the translation of the expert and our system

Google	Apertium	Promt
6%	8%	11%

whether all found words everything can be changed. On it for the long analysis we use "the gold standard" from which could make a start and be sure more precisely above shown table. Rely on 3 experts who know source language well, and the source text is the English context. A task of experts to translate those texts. Sentences translated through experts are compared to offers which have been edited by our system. And as a result coincidence of words between experts and our method has the following values (Table 4).

Considering coincidence we can tell more precisely the system that we offered definitely corrected and improves translation quality.

5 Conclusion

The method consists of two parts: alignment method and method maximum entropy models. This paper focusing on the combination of those two methodology. To find the incorrect word use the method translation memory and on the next stage the maximum entropy method using for editing incorrect words.

As a result proceeding from the received results we can tell what offered post-editing methodic allows to find incorrect word in text and to edit these words. Proposed pos-editing methodic improves quality of machine translated Kazakh text. More exact assessment quality of the translated sentence is planned in the future. It is planned to assess editing the text on quality by using of standard evaluation methods.

References

1. Espla, M., Sanchez-Martinez, F., Forcada, M.L.: 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 (2011)
2. https://en.wikipedia.org/wiki/Principle_of_maximum_entropy
3. Translator Promt. http://www.promt.ru
4. Translator Google. https://translate.google.kz/#kk/en
5. http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/MASHINNI_PEREVOD.html
6. Espla-Gomis, M., Sanchez-Martinez, F., Forcada, M.L.: 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 (2015)

7. Esplà-Gomis, M., Sánchez-Martínez, F., Forcada, M.L.: 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, pp. 172–179
8. Koehn: Statistical Machine Translation, p. 113. http://www.statmt.org/book/
9. http://bazhenov.me/blog/2013/04/23/maximum-entropy-classifier.html
10. Translator Yandex. https://translate.yandex.kz/
11. http://wiki.apertium.org/wiki/Main_Page