Complex technology of machine translation resources extension for the Kazakh language

Diana Rakhimova, Zhandos Zhumanov

Al Farabi Kazakh National University, Laboratory of Intelligent Information Systems, Almaty, Kazakhstan
di.diva@mail.ru, z.zhake@gmail.com

Abstract. The paper is devoted to creating linguistic resources such as parallel corpora, dictionaries and transfer rules for machine translation for low resources languages. We describe the usage of Bitextor tool for mining parallel corpora from online texts, usage of dictionary enrichment methodology so that people without deep linguistic knowledge could improve word dictionaries, and we show how transfer rules for machine translation can be automatically learned from a parallel corpus. All describe methods were applied to Kazakh, Russian and English languages with a task of machine translation between these languages in mind.

Keywords: linguistic resources, low resources languages, parallel corpora, dictionaries, transfer rules.

1 Introduction

Linguistic resources are an important part of any linguistic study. While languages that have been subjects of computational studies for a long time have a lot of resources ready to be used, other languages have an urgent need to develop such resources. Linguistic resources such as monolingual and parallel corpora, electronic dictionaries, and rule dictionaries are very important both for statistical and rule-based language processing. Development of the resources requires a lot of effort and time. It is only logical that low resourced languages need to take all possible opportunities to make that process easier and faster. In this paper, we describe how to use on-line texts and specialized tools to build and improve linguistic resources. With tools and approaches described it is possible to create sufficient amount of different linguistic resources for low resources languages.

Contribution of this work consists of the following: it unites 3 technologies of linguistic resources extension: for parallel corpora, word dictionaries and transfer rules; the combined technology is applied to Kazakh-English and Kazakh-Russian language pairs. Combination of the three technologies allows using their results together for improvement of each other. Larger corpora help to increase coverage of dictionaries. Corpora and dictionaries together help to infer better transfer rules. In common, pro-

adfa, p. 1, 2011.

posed complex (combined) technology of machine translation resources extension allow to improve of machine translation quality.

2 Related works

Creation of linguistic resources that are being considered in this work has been an important task for all the languages. Techniques used in the work have been tried for different language pairs, but not for Kazakh-English or Kazakh-Russian.

Development of parallel corpora using Bitextor has been described in following works. [1] describes Bitextor and apply it for collecting Catalan–Spanish–English parallel corpora. [2] describes creation of English-French parallel corpus. [3] is devoted to Finnish-English parallel corpus. [4] deals with English-Croatian corpus. There are also similar works on Portuguese-English, Portuguese-Spanish, Slovene-English and Serbian-English language pairs.

Dictionary enrichment methodology for people without deep linguistic knowledge is described in [5] for Spanish and in [6] for Croatian. There are no similar works performed for Kazakh or Russian.

Structural transfer rules for Kazakh are described in [7] and [8]. [9] shows an approach to automated generation of structural transfer rule for Spanish-Catalan, English-Spanish, Breton-French language pairs.

The complex technology that is described in the paper has not yet been used as such for one language pair. Only parts of it have been tested and applied to different languages.

3 Building bilingual linguistic corpora using Bitextor

Creation of multilingual parallel corpora is one of the important tasks in the field of machine translation, especially for statistical machine translation. Today, the Internet can be considered a large multi-lingual corpus, because it contains a large number of websites with texts in different languages. Pages of the sites can be considered as parallel texts (bitexts). Bitextor tool is used to collect and align parallel texts from websites.

Bitextor is a free open source application for collection of translation memories from multilingual websites. The application downloads all HTML files from a website, then pre-processes them into a consistent format and applies a set of heuristics to select the file pairs that contain the same text in two different languages (bitexts). Using LibTagAligner library translation memories in TMX format are created from these parallel texts. The library uses HTML-tags and length of the text segments for alignment [1]. After cleaning the resulting translation memory from TMX format tags, we receive a parallel corpus with sentences in different languages aligned with each other.

Previous methods for aligning parallel texts were based on the length of sentences [10]. Systems based on these methods, show good results for languages with a high correlation between the length of the sentences, but their disadvantage is that many texts are not translated sentence-to-sentence. One sentence in the source text may correspond to two or more sentences in the translation. Therefore, Bitextor uses two methods to determine parallel texts: structure of HTML tags and length of sentences.

A key element in Bitextor is the ability to compare file pairs and identify parallel texts in them. To do this, first of all, it uses file metrics (they can be called "fingerprints"), which are determined from numbered text segments. But before comparing file metrics, a set of heuristics is used. After applying heuristics, Bitextor does not need to process every pair of files to compare all of the metrics to each other. Metrics comparison is performed only if the file pair meets all the heuristics. List of heuristics:

1. Comparison of language of the text: if two files are written in the same language, one cannot be a translation of the other.
2. Comparison of file extensions: if within the same site one file is a translation of another file, they usually have the same extension.
3. File size coefficient: this parameter is relative and used to filter a pair of files whose size is different from each other.
4. Total difference between lengths of the texts: this option has the same function as the previous one, but it measures the size of plain text of every file in the symbols.

The process of creating corpora with Bitextor consists of several successive stages described below.

During download stage, website files are copied onto a computer using HTTrack application. This application downloads all HTML files from a multilingual website. Doing that it maintains directory tree structure.

During the next stage, all downloaded files are pre-processed in order to adapt them to the next stages. Bitextor uses LibTidy library to standardize possibly incorrect HTML files into valid XHTML files. It guarantees that tag structure within these files is proper. Original HTML file encoding is converted into UTF-8.

Once the files have been pre-processed, next step is to gather some information needed to compare files and generating the translation memory, such as name and file extension. The language of each text is determined using LibTextCat library. File metrics are also determined on this step.

Information obtained from files is stored in a list, organized in accordance with a position of analyzed file in the directory tree. It makes access to information easier, as file comparison is done level by level.

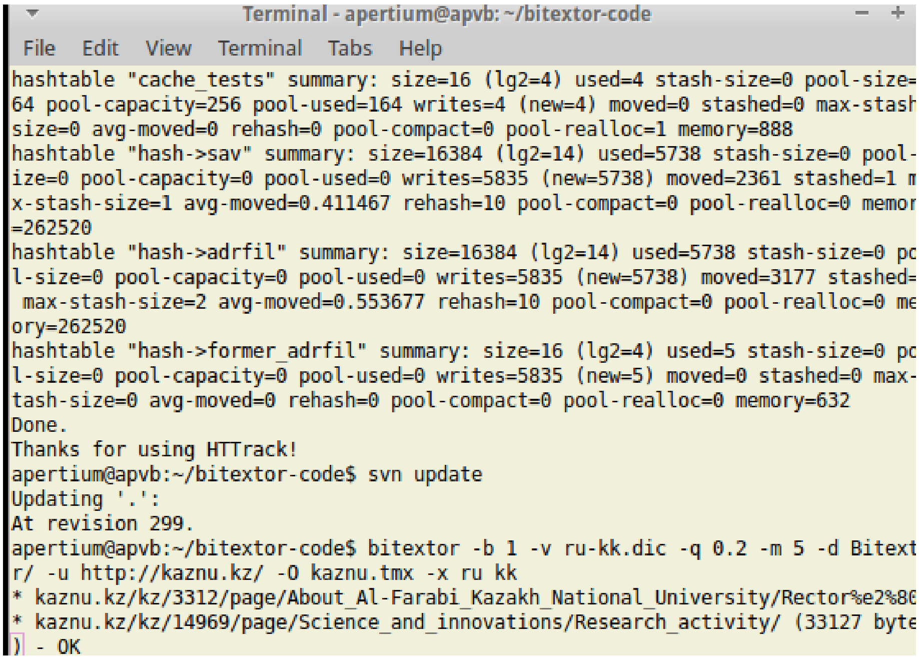
On the stage of comparing files and translation memory generation, comparison of files begins with a comparison of the levels. The user can limit the difference in depth of the directory tree during comparison. Parallel texts, as a rule, are at the same level in the tree or at very close levels, so there is no need to compare each file with all files at different levels.

Generation of translation memory in TMX format is done using LibTagAligner library. As with the metrics, Bitextor uses integer numbers for representing tags and text blocks in TagAligner.

4 Results of the development of bilingual parallel corpora for Kazakh-English and Kazakh-Russian language pairs

Bitextor has beer run for following websites: http://www.kaznu.kz, http://www.bolashak.gov.kz, http://www.enu.kz, http://www.kazpost.kz, http://www.archeolog.kz, http://e-history.kz, http://inform.kz, http://egov.kz, http://primeminister.kz, http://tengrinews.kz and etc.

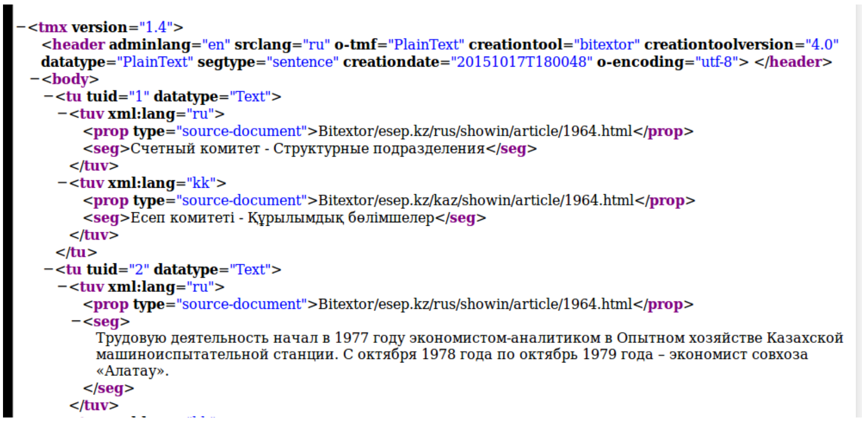
Fig. 1. An example of running Bitextor for www.kaznu.kz



```
hashtable "cache_tests" summary: size=16 (lg2=4) used=4 stash-size=0 pool-size=
64 pool-capacity=256 pool-used=164 writes=4 (new=4) moved=0 stashed=0 max-stash
size=0 avg-moved=0 rehash=0 pool-compact=0 pool-realloc=1 memory=888
hashtable "hash->sav" summary: size=16384 (lg2=14) used=5738 stash-size=0 pool-
ize=0 pool-capacity=0 pool-used=0 writes=5835 (new=5738) moved=2361 stashed=1 m
x-stash-size=1 avg-moved=0.411467 rehash=10 pool-compact=0 pool-realloc=0 memor
=262520
hashtable "hash->adrfil" summary: size=16384 (lg2=14) used=5738 stash-size=0 pc
l-size=0 pool-capacity=0 pool-used=0 writes=5835 (new=5738) moved=3177 stashed=
 max-stash-size=2 avg-moved=0.553677 rehash=10 pool-compact=0 pool-realloc=0 me
ory=262520
hashtable "hash->former_adrfil" summary: size=16 (lg2=4) used=5 stash-size=0 pc
l-size=0 pool-capacity=0 pool-used=0 writes=5835 (new=5) moved=0 stashed=0 max-
tash-size=0 avg-moved=0 rehash=0 pool-compact=0 pool-realloc=0 memory=632
Done.
Thanks for using HTTrack!
apertium@apvb:~/bitextor-code$ svn update
Updating '.':
At revision 299.
apertium@apvb:~/bitextor-code$ bitextor -b 1 -v ru-kk.dic -q 0.2 -m 5 -d Bitext
r/ -u http://kaznu.kz/ -O kaznu.tmx -x ru kk
* kaznu.kz/kz/3312/page/About_Al-Farabi_Kazakh_National_University/Rector%e2%80
* kaznu.kz/kz/14969/page/Science_and_innovations/Research_activity/ (33127 byte
) - OK
```

As a result of Bitextor's work from each site we obtained *.tmx file with the following format:

Fig. 2. A format of obtained parallel corpus for Kazakh-Russian language pair



```
-<tmx version="1.4">
   <header adminlang="en" srclang="ru" o-tmf="PlainText" creationtool="bitextor" creationtoolversion="4.0"
   datatype="PlainText" segtype="sentence" creationdate="20151017T180048" o-encoding="utf-8"> </header>
 -<body>
  -<tu tuid="1" datatype="Text">
   -<tuv xml:lang="ru">
      <prop type="source-document">Bitextor/esep.kz/rus/showin/article/1964.html</prop>
      <seg>Счетный комитет - Структурные подразделения</seg>
    </tuv>
   -<tuv xml:lang="kk">
      <prop type="source-document">Bitextor/esep.kz/kaz/showin/article/1964.html</prop>
      <seg>Есеп комитеті - Құрылымдық бөлімшелер</seg>
    </tuv>
   </tu>
  -<tu tuid="2" datatype="Text">
   -<tuv xml:lang="ru">
      <prop type="source-document">Bitextor/esep.kz/rus/showin/article/1964.html</prop>
     -<seg>
       Трудовую деятельность начал в 1977 году экономистом-аналитиком в Опытном хозяйстве Казахской
       машиноиспытательной станции. С октября 1978 года по октябрь 1979 года – экономист совхоза
       «Алатау».
      </seg>
    </tuv>
```

In this format, tag <tu> includes a pair of aligned segments (in this case - sentences); tag <tuv> - separate sentences in two languages; tag <prop> - HTML file addresses from which these sentences have been extracted; tag <seg> - sentences themselves. In such *.tmx file sentence in one language corresponds to the sentence in another language. It should be noted that comparison quality depends on the website. Thus, we receive a file with parallel texts.

During cleaning of TMX files recurring segments, erroneous and meaningless sentence pairs were deleted. After removal of tags, we received Kazakh-English parallel corpus with 5 925 sentences. This corpus is available at https://drive.google.com/drive/folders/0B3f-xwS1hRdDM2VpZXRVblRRUmM. For Kazakh-Russian language pair we received bilingual parallel corpus of 10 000 sentences.

As it can be seen Bitextor allows saving human and time resources and obtaining parallel aligned corpora from multilingual websites. The corpora then used for ensuring dictionary coverage and for automatic generation of structural rules for machine translation.

5 Automated enrichment of machine translation system dictionaries

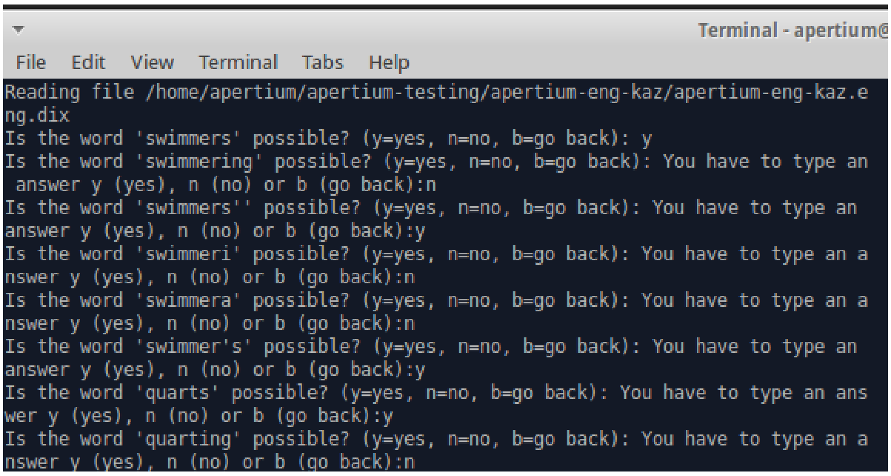
Dictionaries are necessary for translation of texts from one language to another. There are thousands of translation dictionaries between hundreds of languages (English, Russian, Kazakh, German, and etc.) and each of them can contain many thousands of words. Usually, paper version of dictionary is a book of hundreds of pages for which a search for the right word is a fairly long and laborious process. Dictionaries used in machine translation may contain translations into different languages of hundreds of thousands of words and phrases, as well as provide users with additional features.

Such as giving a user an ability to select the languages and translation direction, provide a quick search for words, ability to enter phrases, etc.

Today there are many methods of expanding dictionaries. We used method realized in Apertium by Miquel Esplà-Gomis. We used the tool to fill dictionaries for English-Kazakh, Kazakh-English language pairs in the free/open-source Apertium machine translation system. English-Kazakh MT system has three types of dictionaries: English monolingual, Kazakh monolingual and English-Kazakh bilingual dictionary. All dictionaries, except Kazakh monolingual, have XML format, each word has tag showing which part-of-speech is it [11].

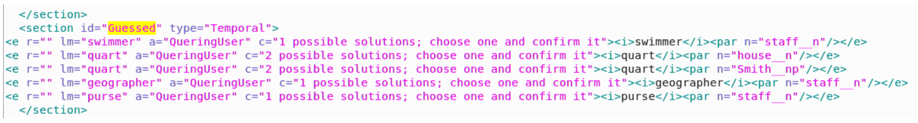
The method is used to assign stems and inflectional paradigms to unknown words if unknown word's paradigm (word pattern) does not appear in dictionaries. The tool needs a file with a list of unknown words that will be added into monolingual dictionary, monolingual dictionary, new dictionary that will be created with the new words added to special section marked “Guessed”, information about a number of questions to be asked. For Kazakh language, it also needs the automation of second-level rules for MT system. The list of unknown words has to be pre-processed with the collection of scripts provided with the tool. After the tool is launched a user can choose among different combination of candidate stems and paradigms correct ones by answering questions asked by system. When a user confirms that the words have been detected correctly they get moved to appropriate dictionary section. In case when the system finds more than one solution for a word all possible options are written to the dictionary along with the number of found possible options.

Fig. 3. Example of using the method for adding words to English monolingual dictionary



```
Reading file /home/apertium/apertium-testing/apertium-eng-kaz/apertium-eng-kaz.e
ng.dix
Is the word 'swimmers' possible? (y=yes, n=no, b=go back): y
Is the word 'swimmering' possible? (y=yes, n=no, b=go back): You have to type an
 answer y (yes), n (no) or b (go back):n
Is the word 'swimmers'' possible? (y=yes, n=no, b=go back): You have to type an
answer y (yes), n (no) or b (go back):y
Is the word 'swimmeri' possible? (y=yes, n=no, b=go back): You have to type an a
nswer y (yes), n (no) or b (go back):n
Is the word 'swimmera' possible? (y=yes, n=no, b=go back): You have to type an a
nswer y (yes), n (no) or b (go back):n
Is the word 'swimmer's' possible? (y=yes, n=no, b=go back): You have to type an
answer y (yes), n (no) or b (go back):y
Is the word 'quarts' possible? (y=yes, n=no, b=go back): You have to type an ans
wer y (yes), n (no) or b (go back):y
Is the word 'quarting' possible? (y=yes, n=no, b=go back): You have to type an a
nswer y (yes), n (no) or b (go back):n
```

Fig. 4. Generated dictionary entries



```
  </section>
  <section id="Guessed" type="Temporal">
<e r="" lm="swimmer" a="QueringUser" c="1 possible solutions; choose one and confirm it"><i>swimmer</i><par n="staff__n"/></e>
<e r="" lm="quart" a="QueringUser" c="2 possible solutions; choose one and confirm it"><i>quart</i><par n="house__n"/></e>
<e r="" lm="quart" a="QueringUser" c="2 possible solutions; choose one and confirm it"><i>quart</i><par n="Smith__np"/></e>
<e r="" lm="geographer" a="QueringUser" c="1 possible solutions; choose one and confirm it"><i>geographer</i><par n="staff__n"/></e>
<e r="" lm="purse" a="QueringUser" c="1 possible solutions; choose one and confirm it"><i>purse</i><par n="staff__n"/></e>
  </section>
```

We have been using this methodology to extend a number of words in dictionaries, which mainly effects to the quality of the translation in machine translation. The technology allows non-expert users who do not have a deep knowledge in computational representation of morphology but understand language being developed participate in building dictionaries. That means more people can add dictionary entries creating larger dictionaries in less time.

6 Automatic generation of structural rules of transformation of sentences

Statistical methods and methods based on rules that are mutually reinforcing approaches to machine translation, which have different strengths and weaknesses. This complementarity appeared as a result growing interest in hybrid systems, combining statistical analysis and linguistic approaches. Therefore, the automatic generation of structural rules based on the transformation of small parallel corpora with further integration in MT system based on rules is a method of helping to solve the above problem in less time and more efficiently. This method avoids the need to manually write these rules of human. To use this method, the program will automatically generate the structural transformation rules for sentences has been adapted for English-Kazakh-Russian and Kazakh language.

Approach by means of which it is possible to receive automatically rules of superficial transfer from small parallel corpora uses the alignment templates (AT) which were originally used in statistical machine translation. The technology of the automated designing of structural rules of machine translation includes a number of stages:

1. Receiving lexical forms by transformation of two parties of the parallel corpus to intermediate representation using the machine translation system Apertium. The intermediate representation consists of lexical forms of words from the case.
2. The lexical form of source language is translated into target language, using the bilingual Apertium dictionary. One word can have several translations in this case rules of grammar of restrictions (Constraint Grammar - CG) [12] or the tagger for parts of speech based on the Hidden Markov Models (HMM), for the solution of a morphological polysemy, and the rule of the lexical choice, for the solution of a lexical polysemy are used.
3. To align, use IBM Model 1.3 and 4 and HMM alignment model for iteration 5 by Giza ++ for two directions of translation. Calculating the alignment Viterbi, according to the model for the two directions of translation. For aligned according to the sentence pairs, two sets of synchronized Viterbi alignments by finding intersections by Och and Ney (2003) [13].
4. The bilingual phrases corresponding to these alignments are taken.
5. For receiving more generalized AT, from bilingual phrases, lexical forms are removed.

6. For finding of optimum AT process of minimization of AT, by way of comparison of attributes and restrictions at the different levels is made, using algorithm of beam search.
7. Further, by use of the list of tags and groups of attributes, the corresponding pairs of languages generation of rules.

Some researchers devote more attention to the alignment step, considering that improving its quality will improve the quality of the entire system.

For the application of the method described above, the program has been adapted for the English-Kazakh-Russian and Kazakh language pair. Create a file with a complete list of classes and attributes of words that describe the morphological characteristics of the lexical forms.

The adapted program has been started on the test corpora of 300 parallel sentences for the Kazakh-Russian and 250 parallel sentences for the English-Kazakh language pair. As a result were received bilingual phrases and structural rules of transformation rules 13 for the Kazakh-Russian and 11 rules for the English-Kazakh language pair.

Table 1. Comparison of methods of transformation rules

Input text	Хорошая школа (Good school)	
	Level of superficial transfer (chunk)	
	Hand-written rules	**The generated rules**
	^adj-noun<NP><sg><p3><PXD><CD>{^жақсы<adj>$^мектеп<n><2><4><5>$}$^sent<SENT>{^.<sent>$}$	^__adj___n_<LRN>{^жақсы<adj>$^мектеп<n>$}$^*executedtule16$^sent<SENT>{^.<sent>$}$
Translation	жақсы мектеп (Good school)	жақсы мектеп (Good school)
Input text	Football player	
	^noun<NP><sg><p3><PXD><CD>{^футбол<n><sg><nom>$^ойыншы<n><2><px3sp><5>$}$^sent<SENT>{^.<sent>$}$	^__n___n_<LRN>{^футбол<n><nom>$^ойыншы<n><px3sp><nom>$}$^*executedtule30$^sent<SENT>{^.<sent>$}$
Translation	футбол ойыншысы (Football player)	футбол ойыншысы (Football player)

As seen in Table 1 of transfers hand-written rules and automatically generated rules are identical.

Automatic generation of structural rules of transformation of sentences is aimed to save time and development effort for creating structural rules. It relies strongly on parallel corpora that are result of using Bitextor as described in sections 3-4.

7 Experiment results

After implementing the technologies describe in the paper experiments on machine translation quality for the language pairs have been conducted. Collected resources were incorporated in Apertium machine translation platform. After that translation quality was compared with Sanasoft and Google Translate – both machine translation applications that support Kazakh-English and Kazakh-Russian language pairs. The results of experiments for Kazakh-English language pairs are shown in tables 2 and 3 .

Table 2. BLEU scores for English-Kazakh translation

MT Application	Estimation %
Google Translate	20,57
Sanasoft	15,74
Apertium	58,97

Table 3. BLEU scores for Kazakh-English translation

MT Application	Estimation %
Google Translate	33,08
Sanasoft	13,97
Apertium	34,5

MT systems have some mistakes in translations. The Google MT system gets a good BLEU score, but in translation of selected phrases, it makes common mistakes such as not assigning the right possessive and case. The Sanasoft system has more errors as regards the translation of words, and many out-of-vocabulary words. Evaluation of English-Kazakh pair machine translation before of extension of linguistic resources was approximately 15%. More biggest evaluation of English-Kazakh after of extension of linguistic resources is explaned by more biggest volume increasing of linguistic resources.

8 Conclusion

In this paper, we describe complex (combined) technology of building linguistic resources for low-resourced languages. We show how to use Bitextor to create parallel corpora, a method of enriching dictionaries with new words without much of linguistic knowledge and how to collect transfer rules for machine translation from a parallel corpus. The results of applying described methods for Kazakh, Russian and English languages show that they allow to save human and time resources, to improve machine translation quality.

References

1. Esplà-Gomis M. Bitextor: a Free/Open-source Software to Harvest Translation Memories from Multilingual Websites. //Proceedings of MT Summit XII, Ottawa, Canada. – Association for Machine Translation in the Americas. – 2009
2. Esplà-Gomis, M. and Forcada, M., 2010. Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with Bitextor. The Prague Bulletin of Mathematical Linguistics, 93, pp.77-86.
3. Rubino, R., Pirinen, T., Espla-Gomis, M., Ljubešic, N., Ortiz Rojas, S., Papavassiliou, V., Prokopidis, P. and Toral, A., 2015, September. Abu-MaTran at WMT 2015 Translation

Task: Morphological Segmentation and Web Crawling. In Proceedings of the Tenth Workshop on Statistical Machine Translation (pp. 184-191).

4. Esplà-Gomis, M., Klubicka, F., Ljubesic, N., Ortiz-Rojas, S., Papavassiliou, V. and Prokopidis, P., 2014, May. Comparing two acquisition systems for automatically building an English-Croatian parallel corpus from multilingual websites. In LREC (pp. 1252-1258).
5. Espla-Gomis, M., Carrasco, R.C., Sánchez-Cartagena, V.M., Forcada, M.L., Sánchez-Martınez, F. and Pérez-Ortiz, J.A., 2014. An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words. In Proceedings of the 17th Annual Conference of the European Association for Machine Translation (pp. 19-26).
6. Ljubešic, N., Espla-Gomis, M., Klubicka, F. and Preradovic, N.M., Predicting Inflectional Paradigms and Lemmata of Unknown Words for Semi-automatic Expansion of Morphological Lexicons. RECENT ADVANCES IN, p.379.
7. Sundetova, A., Karibayeva, A. and Tukeyev, U., 2014. Structural transfer rules for kazkah-to-english machine translation in the free/open-source platform Apertium. TÜRKİYE BİLİŞİM VAKFI BİLGİSAYAR BİLİMLERİ ve MÜHENDİSLİĞİ DERGİSİ, 7(1 (Basılı 8).
8. Sundetova A., Forcada, M.L., Shormakova, A. & Aitkulova, A. (2013). Structural Transfer Rules for Kazakh-to-English Machine Translation In the Free/OpenSource Platform Apertium. In Proceedings Of the I International Conference on Computer processing of Turkic Languages (TurkLang'13), p. 322-331. Astana, Kazakhstan.
9. Sánchez-Cartagenaa V.M., Pérez-Ortiza J.A., Sánchez-Martínez F. A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. //Computer Speech & Language. - July 2015. - Volume 32, Issue 1. – P.46–90.
10. Brown P., Lai J., Mercer R. Aligning sentences in parallel corpora. //Proceedings of the 29th annual meeting on Association for Computational Linguistics. – Association for Computational Linguistics Morristown, NJ, USA. - 1991. – P.169–176.
11. Forcada M. L., GinestíRosell M., Nordfalk J., O'Regan J., OrtizRojas S., PérezOrtiz J. A., SánchezMartínez F., RamírezSánchez G., Tyers F. M. Apertium: a free/opensource platform for rulebased machine translation. //Machine translation. - 2011. - 25(2). - P.127144.
12. Karlsson F., Voutilainen A., Heikkilä J., Anttila, A. Constraint Grammar: A language independent system for parsing unrestricted text. //Mouton de Gruyter. – 1995.
13. Och F.J., Ney H. A systematic comparison of various statistical alignment models. //Computational Linguistics. – 2003. - 29(1). – P.19–51.