Machine Translation Infrastructure for Turkic Languages (MT-Turk)

Emel Alkım and Yalçın Çebi
Department of Computer Engineering, Dokuz Eylul University, Turkey

Abstract: *In this study, a multilingual, extensible machine translation infrastructure for grammatically similar Turkic languages "MT-Turk" is presented. MT-Turk infrastructure has multi-word support and is designed using a combined rule-based translation approach thatunites the strengths of interlingual and transfer approaches. This resulted in achieving ease of extensibility by adding new Turkic languages. The new language can be used both as destination and as source language achieving two-way extensibility. In addition, the infrastructure is strengthened with the ability of learning from previous translations and using the suggestions of previous users for disambiguation. Finally, the success of MT-Turk for three Turkic languages -Turkish, Kirghiz and Kazan- is evaluated using BiLingual Evaluation Understudy (BLEU) metric and it is seen that the suggestion system improved the success by 43.66% in average. Although the lack of linguistic resources affected the success of the system negatively, this study led to the introduction of an extensible infrastructure that can learn from previous translations.*

Keywords: *Rule-based machine translation, Turkic languages, semi-language specific interlingua and disambiguation by suggestions.*

Received April 21, 2015; accepted November 8, 2016

1. Introduction

Communication has become the most crucial topic in the era of globalization and internet. However, existence of different languages forms a barrier to connectivity. Machine Translation (MT) is the main method to achieve this connectivity and overcome the language barrier. With the amount of parallel corpora increasing and the difficulty of constructing a rule-based high quality MT, the research has dominated towards corpus-based approaches [14]. However, Turkic languages are under resourced languages and there are no parallel corpora available to construct corpus-based machine translation tools. Fortunately, they are grammatically similar as they come from the same language family and machine translation is easier and more applicable for grammatically similar languages [13], especially as they show similar structural and semantic properties [5].

Machine translation between Turkic languages has been addressed using different methodologies on different language pairs. Apertium Turkic working group is a part of Apertium project, an open source platform for rule-based machine translation. In Apertium, lexical processing is achieved by finite-state transducers, whereas hidden Markov models are used for part-of-speech tagging, and multi-stage finite-state chunking is used for structural transfer [12]. Each language pair is added to the system separately and they have various language pairs in different levels of quality [6]. However, the only pair which is reported as release quality is Kazakh-Tatar language pair [19]. Tantug proposed a hybrid system which combines rule-based and statistical approaches using two level morphology [22]. A Turkmen to Turkish translation system [24] and a Uyghur to Turkish translation system [17] are developed using that study and it is stated that addition of new languages is possible for translation from any Turkic language to Turkish [22]. However, it is not possible to translate from Turkish to the new language as they use Turkish corpus for disambiguation. Dilmaç is also a multilingual project [11] which is started by Turkmen-Turkish morphological analysers and translator [20]. The translation is performed word by word, it does not support multi-words and the lexicon size is very limited for some languages. Tayirova *et al.* [26] applied n-gram based and phrase based statistical machine translation on Kirghiz-Turkish language pair [26].

In this research, a rule-based Machine Translation Infrastructure for Turkic languages (MT-Turk) was developed to address the need of a multilingual, two-way extensible machine translation infrastructure with multi-word support. MT-Turk is designed in a rule-based and multilingual manner so that new languages can be added by supplying necessary information; particularly the lexicon, morphological rules, phonological rules and suffixes of that language. The new language can be used as destination as well as source, providing a two-way extensible structure.

In order to evaluate the quality of the machine translation system, MT-Turk is initially developed and tested for Turkish and Kirghiz. Then, the results and the problems were reported in Turkic Language conferences to discuss the issues and possible resolutions [3, 4]. Subsequently, Kazan Tatar is added to the system as the third language.

In the remainder of the article the problem domain, Turkic languages, is defined briefly. Then, the infrastructure is described in detail. Finally, the evaluation sets and results are discussed before concluding with summarization of the main outcomes.

2. Turkic Languages

Turkic language family, which belongs to the Ural-Altaic group [21], consists of 40 languages which are closely related to each other.

Grammatically, the most significant property of Turkic languages is that they are agglutinative languages in which the words are formed by adding affixes to root. Therefore, a single word can represent a whole sentence and morpho-syntactical information is very important for analysing and translating the text. Such an example where a Turkish word forms an English sentence with thirteen words is:

- Çekoslavakyalılaştıramadıklarımızdansınız.
- English: You are one of those that we could not turn into a Checkoslavakian.

Turkic languages are relatively similar however minor differences in the structures of the languages make translation harder. Translations of two sentences in eight Turkic languages are listed in Table 1.

Most significant differences that affect the translation performance are problems with suffix correspondences and suffix binding changes. One significant problem is that a suffix in one language can be translated to the other language using a suffix group instead of one corresponding suffix. Such an example is one of the past tenses in Kirghiz “GAn”. The Turkish translation of this suffix is achieved by using two suffixes together, Past Tense “mIş” and aspect “DI”.

Additionally, although word order is the same in all Turkic languages Subject Object Verb (SOV), suffixes can require binding to different members of a phrase in some languages. For example, the Kirghiz phrase “casagan işter*in*: *the jobs he did*” is formed by adding a participle to the verb and a possessive suffix the noun in the structure “verb+participle noun+plural+ possessive”. However in Turkish the possessive suffix should be added to the end of the participle as “yaptığ*ı* işleri” forming the structure “verb+participle+possessive noun+plural”.

In MT-Turk, the richness of the morpho-syntactical information is addressed by a rich analyse phase and the differences between the structures of Turkic languages are addressed by transfer phase in terms of both the stems and phrases. The details of the phases are described in the following section.

3. MT-Turk

In this study, an extensible and self-extending machine translation infrastructure for Turkic languages was developed in a rule-based manner. Two subsets of rule-based approach, the interlingual machine translation approach [9] and transfer-based approach [15], were used in combination to form the multilingual machine translation system to achieve extensibility and interoperability. The input is analysed to form a semi-interlingual representation and then this semi-interlingual form is transferred to the target language’s semi-interlingual form using transfer rules.

The stems are translated using the interlingual machine translation approach. The sentences in source language are analysed and each word or multi-word group is converted to a language-neutral representation of the concept they identify. A concept in MT-Turk is common to more than one language and thus, no bilingual transfer dictionaries are required. As a result, the most crucial and problematic resource of the translation system, lexicon, can be enhanced easily. On the contrary, the suffixes and word order changes are achieved using transfer based machine translation approach.

Extensibility of the system is achieved by the semi-interlingual representation. There is no need for specific analysers and generators for language-pairs. Each new language only needs to specify its rules and lexicon so that it can be added to MT-Turk.

Disambiguation and forming a fully language independent canonical representation to construct pure interlingua is very difficult for Turkic languages as a result of their under resourced property (lack of a large corpus). However as Turkic languages are closely related; structural and semantic properties are similar and the semi-interlingual representation is sufficient with an additional transfer phase. The main architecture of the system is shown in Figure 1.

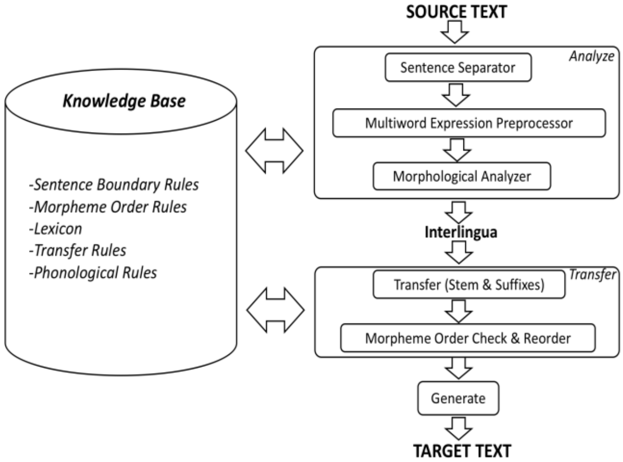


Figure 1. The architecture.

The components of the knowledge base and the translation process are discussed in the remainder of this section.

Table 1. Translation of two sentences in eight turkic languages [10].

Language	Sentence 1	Sentence 2
English	Who telephoned dad?	He drank three glasses of tea.
Turkish	Kim telefon etti baba?	Üç bardak çay içti.
Azerbaijani	Zängedän kim idi, ata?	Üç stäkan çay içdi.
Baskurt	Kim telefon itti, atay?	Öss takan säy isti.
Kazakh	Kim telefon soktı, ata?	Üş kese (stakan) şay işti.
Kirghiz	Kim telefon çaldı, ata?	Üç çını çay içti.
Uzbek	Kim ḳonğıràḳ ḳıldi, dädä?	Üç stäkän çày içti.
Tatar	Kim telefon etti, äti?	Öç stakan çäy içti.
Turkmen	Kim telefon etdi, kāka?	Üç stakan çāy içdi.
Uyghur	Kim telefon kıldi, dada?	Üç çinäsiŋ çay içti.

3.1. Knowledge Base

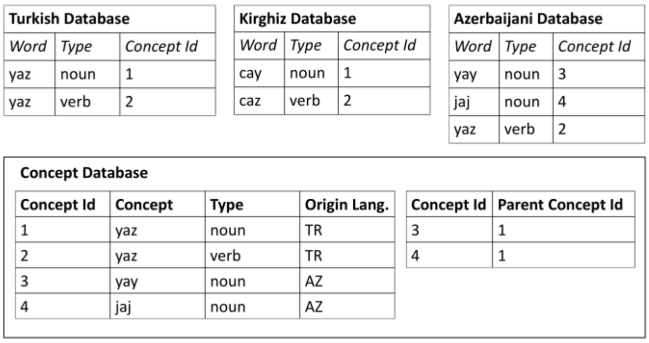
The knowledge base consists of the rule files and the lexicon that are required during the translation process. The main knowledge components are:

- *Sentence boundary rules*: are stored in XML format and used by the Sentence Separator component to detect sentence boundaries.
- *Morpheme order rules*: are used by morphological analyzer to check morpheme order. The validity of the morpheme order is checked by using three rule files: "morpheme ordering rules," "must rules" and "not rules" [7]. "Morpheme ordering rules" lists all of the possible morpheme sequences that can result in a valid word. "Must rules" are used to define constraints that must be achieved, more specifically which suffix should precede the other if they exist in the same word. "Not rules" on the other hand specify the tag sequences that must be avoided, i.e., the suffixes that cannot occur in the same word.
- *Lexicon*: Some multi-lingual applications such as MulTra project [28] use bilingual lexicons for each language pair in the system. Although bilingual lexicons are efficient for holding word-pairs, a new problem arises: for a multilingual system that has already three languages, adding a new language to the system requires constructing three separate bilingual lexicons, one for each language that is already in the system. Thus, the enhancement of the system is hard and has a high space and time cost.

As the main aim of this study is to achieve an extensible system that has no pivot language, each language has its own database. Each database consists of stems, suffixes and their alternations.

Although separate databases are used, a connection between the databases has to be provided to achieve transfer from one language to another. However, the stems of two languages are not connected in a one-to-one manner. In every language, there are words that express more than one meaning with a single representation. For example, "yaz" in Turkish is used both as a noun with meaning "summer season" and also a verb with meaning "write". These meanings can have different representations in another language. Moreover, for some other language, some of these meanings can have more than one different representations.

In MT-Turk, a common database of concepts is used to achieve the connection between languages. This common database is responsible for holding the list of concepts introduced to the system so far by the lexicon of each language. A concept in MT-Turk is a notion that corresponds to an entity that has a specific representation and one or more meanings. For example, if in language A, there is one representation for two words (synonyms) in language B; then there is a general concept for the meaning in language A whereas there are two in language B. Therefore, each concept must be stored in the common database. The list of the concepts can contain entries from different languages, i.e., one concept can be introduced by Turkish, another by Kirghiz or Kazan Tatar. The common database also contains the covering relations of concepts. The contents of these tables for the example stem "yaz" are listed in Figure 2.



Turkish Database

Word	Type	Concept Id
yaz	noun	1
yaz	verb	2

Kirghiz Database

Word	Type	Concept Id
cay	noun	1
caz	verb	2

Azerbaijani Database

Word	Type	Concept Id
yay	noun	3
jaj	noun	4
yaz	verb	2

Concept Database

Concept Id	Concept	Type	Origin Lang.
1	yaz	noun	TR
2	yaz	verb	TR
3	yay	noun	AZ
4	jaj	noun	AZ

Concept Id	Parent Concept Id
3	1
4	1

Figure 2. Sample representation of the relation between CONCEPTSET and languages.

Algorithm 1 shows the lexicon storage infrastructure. For each new input language, the algorithm creates a mapping between each representation to the corresponding translations of it in the languages that are already in the system.

Algorithm 1: Lexicon storage and concepts' set construction algorithm.

```
Input:  U – a list of languages currently in MT-Turk
        l – input language.

For each language l' in U
    create a mapping between each representation w in
    language l and representation w′ in language l′
        if $l(w) \rightarrow l'(w')$ then
            SET_CONCEPT($w, l'(w')$)
if $l(w) \rightarrow l'(w_1') \& l'(w_2')$ then
            new_concept=STORE_CONCEPT(w,l)
            SET_PARENT($l'(w_1')$, new_concept)
            SET_PARENT($l'(w_2')$, new_concept)
        if $l(w_1) \& l(w_2) \rightarrow l'(w')$ then
            new_concept1=STORE_CONCEPT($w_1$, l)
            new_concept2=STORE_CONCEPT($w_2$, l)
```

```
12:        SET_PARENT(new_concept1,$l'(w')$)
13:        SET_PARENT(new_concept2,$l'(w')$)
```

If there is a one-to-one mapping, a new concept is not created. The new stem is stored with the concept information of the existing concept. If one stem in the new language corresponds to two or more representations in the existing language, a new concept originated from the single representation in the new language is stored in the concepts table and this concept is stored as the parent of the corresponding concepts of the existing language. If more than one representation in the new language corresponds to one representation in the existing language, new concepts originated from these representations are stored in the concepts table and the concept of the existing language's representation is stored as the parent of these concepts.

- *Transfer rules*: are used at the transfer phase. In some situations one tag in one language has to be represented with a tag sequence in the second language. For example, a past tense suffix in Kirghiz "Type Two Past Tense GAn/GOn" must be translated to Turkish with a sequence of past tense "mIş" and a copula "DI", as the correct translation for "baştaganmın - I had begun" is "başlamıştım". Hence, each suffix tag in a language must have either a corresponding tag or a tag sequence in each language that is defined in the system and there must be a mapping of this correspondence with transfer rules. During translation, these rules are used to transform the analyzed semi-interlingua to the correct form by replacing the tag with the tag sequence or vice versa.
- *Phonological rules*: are used for the analysis, alternation and the generation purposes. The substitutions are run at the first phase of generation process and they are used for character substitutions in suffix representations like A → a | e. The rules, on the other hand, are used for defining constraints on how phonemes can be used together. Both substitutions and rules consist of two parts: match and action. Match part is used to decide if rule should be applied whereas action part defines the action to undertake (like replacing a character with its allomorph). The alternations for each stem and suffix are generated by the phonology library using phonological rules and then used by the analyzer.

3.2. Translation Process

MT-Turk translates the input to the target language in three phases: analyse, transfer and generate. At the analysis phase, firstly the input is separated to sentences using the sentence separator developed by Aktaş [2]. Then, the multi-word expressions are extracted and each word or multi-word is analyzed morphologically using aparametric version of the morphological analyser developed by Birant [7].

Multi-Word Expressions (MWE) are defined as structures with more than one word, whose structure and meaning cannot be derived from their component words' independent meanings [27]. MWE is a very complicated and problematic issue for natural language processing applications especially for morphologically rich languages like Turkish.

The MWE in Turkish can be grouped under four types [16], which are:

- *Lexicalized collocations*: MWEs that are formed with duplication of same word or in a predefined structure. (e.g.: *hiç olmazsa*: at least; *ipe sapa gelmez*: nonsensical).
- *Semi-lexicalized collocations*: MWEs that are already stored in the database (e.g.: *kafayı ye-*: go nuts).
- *Non-lexicalized collocations*: MWEs that are formed with use of some suffixes (e.g.: *koş-a koş-a*: by running; *uyu-r uyu-maz*: as soon as he sleeps).
- *Multi-word Named-entities:* MWEs that are proper names (e.g.,: *Dokuz Eylül Üniversitesi*: Dokuz Eylul University).

Each type needs special attention and different strategy. The first two types of MWEs are the ones that are matched from the lexicon. The multi-word expressions of these kinds are contained in the existing Turkish lexicon as they represent a different meaning from the independent meanings of its component words' meanings. The input text is analysed by the multi-word expression pre-processor prior to the morphological analysis. The pre-process is done by gathering possible multi-word list from the lexicon and searching the input text for the existence of these multi-words. The multi-word groups of the third type are handled by multi-word rules that are specific morphophonemic rules. The multi-word rules are specified with a special tag <mwrule> (Multi-Word Rule). Each rule should specify the group name, the lexical form and the surface form of the match structure. The match structure can define structures to be matched in more than one adjacent word with different suffixes to be matched in each word. Some special abbreviations are used by the match structure: W for identifying a word followed by the index (order) number of the word and # is used to identify word boundary. A sample multi-word group construction rule ("ir_mez", as in "gelirgelmez: as soon as he comes") is shown in Figure 3. It is specified in the lexical form that first word must have the suffix "YtuU1 -Ir" followed by the word boundary and the second word must have the suffix "YtuU2-mAz". The surface form of this group is formed by enclosing the two matched words with a group tag.

```
<mwrule>
<group>ir_mez</group>
<lex>
        W1
<YtuU1>Ir</YtuU1>
        W2
<YtuU2>mAz</YtuU2>
</lex>
<surf>
<group name="ir_mez">W1#W2</group>
</surf>
</mwrule>
```

Figure 3. An examplemulti-word rule “ir_mez”.

During the multi-word pre-process, the input string is analysed and all the multi-word rules are checked to see if the lexical structure of the rule is matched. If it is matched, it is applied by transforming the matched structure to form the surface structure. A sample interlingua that is formed by the multi-word rule application in the transfer phase is shown in Figure 4.

```
<Group name="ir_mez">
<Word>
<ValueOfWord> gelir </ValueOfWord>
<Root index="2545">
<Value> gel </Value>
</Root>
<Suffixes>
<SuffixCombination index="0">
<YtuU1> ir </YtuU1>
</SuffixCombination>
</Suffixes>
</Word>
<Word>
<ValueOfWord> gelmez </ValueOfWord>
<Root index="2545">
<Value> gel </Value>
</Root>
<Suffixes>
<SuffixCombination index="0">
<YtuU2> mez </YtuU2>
</SuffixCombination>
</Suffixes>
</Word>
</Group>
```

Figure 4. Sample group interlingua.

In MT-Turk a modified interlingua approach is used due to the output of the morphological analyser. The output of the original morphological analyser is a list of all possible root-suffix combinations in Extensible Markup Language (XML) format and is language dependent; i.e., it contains the value of the root and the suffixes in the source language. This structure forms an output XML that is easy to read and interpret even by the human eye.

However, the interlingua must be language independent so that it can be used during translation between any two languages. The language independency is achieved by adding concept id to the root information in the output of the morphological analyser. The concept id is hold in the Index attribute of the Root tag and the morphological analyser must be called with a specific parameter to return the output in this format. The use of the concept id achieves language independency in stems as it is common to all languages. Suffix tags are also common between languages but a transfer mechanism for the suffix tags is required during translation for handling exceptions.

Consequently, although the roots are language-independent, the tags are still in the source language and also the values of the roots and suffixes are still present. Hence the output is semi-language specific, i.e., the interlingua has the language specific output of the analyser with the concept information of the roots added.

The main intention behind keeping the original design with a small addition is to maintain high readability of the XML output in addition to maintain interoperability between the morphological analyser and previously developed tools for Turkish. High readability of the XML output is very useful especially during language resources development process. A sample semi-language specific interlingua on the word “*evde* (at home)”, which is the output of morphological analysis, is given in Figure 5.

```
<Word>
<ValueOfWord> evde </ValueOfWord>
<Root index="25313">
                    <Value>ev</Value>
</Root>
<Suffixes>
<SuffixCombination index="0">
<DuADurBul>de</DuADurBul>
</SuffixCombination>
</Suffixes>
</Word>
```

Figure 5. Semi-language specific interlingua sample.

As a result of a semi-language specific interlingua, the output of the analyser needs an additional transfer process before it can be generated in the destination language. The stem from the source language should be replaced by the corresponding stem in the destination language and the required transformations for suffixes must be achieved.

Translating the word “yay-*summer*” from Azerbaijani to Turkish is done by following the steps below:

- Get the concept id of “yay” → 3.
- Search stem with concept id 3 in target database (TR: Turkish) → concept id 3 is not found.
- Search for the parent of the concept → parent of concept 3 is concept 1.
- Search parent concept in target database (TR) → “yaz”-noun.

Then the suffix transfer is achieved using the transfer rules. For example; translating the suffixes of the word “baştaganmın - I had begun” from Kirghiz to Turkish is done by following the steps below:

- Get the suffix combination of “baştaganmın” →ganmın : DuEZG2+DuEKGr2T1.
- Check for suffix combinations to be replaced → no matching combination for DuEZG2+DuEKGr1.
- Find correspondences for the existing suffixes→ replace DuEZG2 with DuEZGM (past tense “mIş”) +KDi (copula “DI”).
- The new form of the suffix combination is DuEZGM+KDi+DuEKGr2T1.

At the final step of transfer phase, the morpheme reordering is achieved by the transfer component. Firstly, all the coalescence consonants are removed from the analyser’s output. Then, the validity of this form is checked through the morpheme ordering rules in the destination language. If it is not valid, the morpheme sequence is reordered in a combinational manner and all combinations are checked through the morpheme ordering rules. The processes of morpheme reordering and checking the validity of the new order in the transfer phase are executed in parallel for speeding up the translation time.

Finally, the output in the target language is generated by adding the coalescence consonants if necessary and selecting the correct allomorphs of the characters according to phonological constraints.

In MT-Turk, a suggestion system is integrated in the system to assist in disambiguation and enhancing the success of the system. The suggestions are collected from the users and stored with context information (i.e., the sentence it was used in). Therefore during a new translation; when there is an ambiguity in a word that was suggested before, earlier suggestions are shown to the user in a descending order of the suggestion counts. Consequently, disambiguation is achieved by human interaction who is the native speaker of target language.

4. Discussion

MT-Turk is tested on three Turkic languages: Turkish, Kirghiz and Kazan Tatar. The evaluation is carried out using bilingual texts from Kirghiz to Turkish and Kazan Tatar to Turkish with two reference translations. One of the translations is used as the source and the original text is used as the reference translation. The Kirghiz-Turkish evaluation set contains 263 sentences whereas Kazan Tatar-Turkish evaluation set contains 127 sentences. Due to lack of parallel corpora available, translations of a Turkish text to Kirghiz and Kazan Tatar is used for the evaluation of translation between Kirghiz and Kazan Tatar. Unfortunately, Kirghiz-Kazan Tatar evaluation set contains only 18 sentences. Kazan Tatar is introduced to the system to evaluate MT-Turk’s extensibility in addition of a new language. However, the lack of parallel lexicons and data sets, which is the main motivation behind using a rule-based design, also affected the evaluation process and caused the evaluation sets to be very small, especially in Kirgiz-Kazan Tatar language pair.

In the case of an ambiguous translation, the topmost ambiguity is chosen as the translation candidate. Therefore, if there are prior suggestions, the one with the highest score is chosen.

The evaluation is achieved using BiLingual Evaluation Understudy (BLEU) metric [18]. BLEU is a metric which is independent from the source language, however it is not sufficient enough to be used as a comparison technique between machine translation systems [29] and not efficient especially on agglutinative languages as a mistranslated suffix can produce a total mismatch [24]. Moreover, it is more extensively used in evaluation of corpus-based systems not rule-based systems. In this study BLEU used as a way to compare the successes of different languages in the system and also to measure the effect of suggestion over translation success.

The achieved BLEU scores with and without suggestions are listed in Table 2. The BLEU scores are lower from Turkish to Kirghiz and Turkish to Kazan Tatar due to higher number of lexicon in Turkish and as the additional lexicon was formed by the correspondences of Kirghiz and Kazan Tatar words, not focusing on a full lexicon. Furthermore, the BLEU score is also affected negatively as there is only one evaluation reference.

The success of translation between Turkish and Kazan Tatar is lower as Kazan Tatar has fewer resources in the system. However the success of translation between Kirghiz and Kazan Tatar is higher as they are closer languages.

It can also be seen from Table 2 that suggestion improves the success of the translation between 34.77% and 56.20% with an average of 43.66%. The minimum improvement is achieved on translation from Kirghiz to Kazan Tatar whereas maximum is achieved on translation from Kazan Tatar to Turkish.

A sample Kirghiz sentence and the translation output in Turkish are given in Table 3 together with the two reference sentences to analyse the failures of the translation further. When the outputs of the translation is studied, it is seen that the first word “kapçıgay” were translated with the correct stem butmissing a suffix, the reason for this is that there is no “kapçı” or “kapçıgay” in any of the dictionaries but the grammar book [8] where this sentence and its translation is retrieved contains the word “kapçıgay” and is translated as “kanyon” (canyon), thus “kapçıgay” is stored in the lexicon as “kanyon”. The second, third, fourth and fifth words were translated correctly. The sixth word is translated with a different suffix and although both of the translators used -ten, the suffix correspondent of -dı (accusative) in Kirghiz is listed as -I in Turkish by the grammar books.

The word group “aylana berip” is translated as “dönüverip” because “ber” is the auxiliary verb that has the correspondent “-iver” in Turkish. The last two words are translated correctly. As a result, the analysis shows that failures are mostly caused by the lack of lexicon.

Table 2. Evaluation results (BLEU).

	Without Suggestion	**With Suggestion**	**Suggestion Improvement**
Kirghiz→Turkish	15.12	21.71	***43.58%***
Turkish →Kirghiz	8.65	12.34	***42.66%***
Kazan Tatar→Turkish	9.52	14.87	***56.20%***
Turkish →Kazan Tatar	5.04	7.20	***42.86%***
Kirghiz→Kazan Tatar	13.46	18.14	***34.77%***
Kazan Tatar→Kirghiz	14.23	20.19	***41.88%***
Average	*11.00*	*15.74*	***43.66%***

Table 3. Kirghiz sentence translation output.

Kapçıgay ördögön cüktüü maşina taş moynoktu aylana berip tık toktodu. *The loaded car that moves at the canton stopped just after turnıng the rocky bend.*	Kirghiz Text
Kanyonda ilerleyen yük arabası Taş Moynok tan döner dönmez hemen durdu.	Reference 1
Dağ geçidine doğru ilerleyen yüklü araba, taşlı dönemeçten geçerken tık durdu.	Reference 2
Kanyon ilerledikçe yüklü araç taş dönemeci dönüverip tık tavuktaydı.	Output (No suggestion)
Kanyon ilerleyen yüklü araba taş dönemeci dönüverip tık durdu.	Output (Withsuggestion)

There are not many reported BLEU scores for machine translation systems between Turkic languages. An English-to-Turkish statistical machine translation system that is evaluated on 649 sentences achieved a BLEU score of 27.64 [25]. Whereas another study on Turkish, a Turkmen-to-Turkish machine translation system, which is evaluated on 254 sentences, achieved a BLEU score of 33 [23]. Tayirova *et al.* [26] evaluated the system they developed on Kirghiz-Turkish language pair using 100 short and 100 long sentences in both n-gram based and phrase based statistical machine translation and reported the average BLEU score as 10. Two first two studies which only evaluates one way (translations to Turkish) have higher scores, whereas the statistical approach by Tayirova *et al.* [26] has the lowest score.

5. Conclusions

This paper presents a rule-based MT-Turk for Turkic Languages. In MT-Turk, all possible translations are listed instead of choosing one, as most powerful disambiguation techniques require a corpus, which reveals a problem for low resource languages like Kirghiz and Kazan Tatar. Disambiguation is left to humanmind that is the ultimate disambiguator. The system is empowered by a suggestion system, in other words, the ability to learn how to disambiguate from the user. The translation with the highest suggestion number is listed at the top of the possible translations.

The final BLEU score of MT-Turk changes between 7.20 and 21.71 at different language pairs and directions. The highest score is achieved by Kirghiz to Turkish translation as Turkish and Kirghiz are the languages with most linguistic resources. The worst score is the evaluation result of Turkish to Kazan Tatar translation as Kazan Tatar has very few resources and Turkish has many. The average BLEU scores without suggestion and with suggestion are 11 and 15.74 respectively.

The suggestion system improves the success of the translation 43.66% in average. For the purposes of checking the integrity of the BLEU scores, the Turkish translations of the original Kirghiz text were evaluated with reference to each other, selecting the second reference as the candidate translation and the BLEU score was evaluated as 10.31. This is an indication of how poor BLEU metric performs on agglutinative Turkic languages on the basis of Kirghiz-Turkish language pair.

MT-Turk provides a complete rule-based infrastructure for machine translation between Turkic languages, therefore; adding a new Turkic language can be achieved by just adding the dictionary of stems, suffixes and the rules. There is no need for bilingual dictionaries and the new language can be used as either destination or source; therefore, MT-Turk is two-way extensible. MT-Turk has support for multi-word structures and “*suffix-to-suffix combination*” suffix correspondences. Furthermore, MT-Turk keeps learning from users’ suggestions and improving the translation quality. Consequently, the scope and extensibility of MT-Turk will help improving the unity of Turkic communities on written work of art and obtain fusion of Turkic communities.

Goals and directions for future work includes extending the MT-Turk infrastructure by adding other Turkic languages, extending the resources of the existing languages for better performance and extending the evaluation sets to make the evaluation more effective. Future research can also include developing a Cross Language Information Retrieval (CLIR) tool for Turkic languages on top of MT-Turk architecture. There are a number of CLIR tools available in the literature [1]; however, there is none available for Turkic languages.

This work is partly supported by Dokuz Eylül University Scientific Research Projects Coordination Unit under the contract number 2009_KB_FEN_11 and is a product of the Ph.D. thesis at the Graduate School of Natural and Applied Sciences of Dokuz Eylül University.

MT-Turk can be accessed at the applications site of DokuzEylul University Natural Language Processing Research group http://nlp.cs.deu.edu.tr. A direct link to MT-Turk is http://nlpapps.cs.deu.edu.tr/MTTurk/.

Some functionalities and tools, like suggestion, can be accessed by members only. An account can be requested by contacting the group or the authors.

References

[1] Ahmed F. and Nurnberger A., "Literature Review of Interactive Cross Language Information Retrieval Tools," *The International Arab Journal of Information Technology*, vol. 9, no. 5, pp. 479-486, 2012.

[2] Aktaş Ö., "Türkçe için Verimli bir Cümle Sonu Belirleme Yöntemi," *in Proceedings of Akademik Bilişim Bilgi Teknolojileri Kongresi IV*, Denizli, 2006.

[3] Alkım E. and Çebi Y., "Türk Lehçeleri Arası Otomatik Çeviri ve Karşılaşılan Sorunlar," *in Proceedings of V. Genç Türkologlar Sempozyumu Kitabı*, Bishkek, 2012.

[4] Alkım E. and Çebi Y., "Türk Dillerinin Bilgisayarlı Çevirisi ve Karşılaşılan Sorunlar," *in Proceedings of VII Uluslararasi Turk Dili Kurultayi*, Ankara, 2012.

[5] Altıntaş K. and Çiçekli İ., "A Machine Translation System Between a Pair of Closely Related Languages," *in Proceedings of the 17th International Symposium on Computer and Information Sciences*, Orlando, pp. 192-196, 2002.

[6] Apertium, "Apertium Turkic Working Group," 2016. [Online]. Available: http://wiki.apertium.org/wiki/Apertium_Turkic, Last Visited, 2016.

[7] Birant, C., "Root-Suffix Seperation of Turkish Words," Thesis, Dokuz Eylül Üniversitesi, 2008.

[8] Çengel H., *Kırgız Türkçesi Grameri-Ses ve Şekil Bilgisi*, Akçağ Yayınları, 2005.

[9] Dorr B., Hovy E., and Levin L., *in Encyclopedia of Language and Linguistics*, Elsevier, 2006.

[10] Ercilasun A., *Karşılaştırmalı Türk Lehçeleri Sözlüğü*, Kültür Bakanlığı Yayınları, 1992.

[11] Fatih University, "DİLMAÇ Project," 2013. [Online]. Available: http://datamining.ceng.fatih.edu.tr:8080/dilmac/, Last Visited, 2013.

[12] Forcada M., Ginestí-Rosell M., Nordfalk J., O'Regan J., Ortiz-Rojas S., Pérez-Ortiz J., Sánchez-Martínez F., Ramírez-Sánchez G., and Tyers F., "Apertium: A Free/Open-Source Platform for Rule-Based Machine Translation," *Machine Translation*, vol. 25, no. 2, pp. 127-144, 2011.

[13] Hajič J., Hric J., and Kubon V., "Machine Translation of Very Close Languages," *in Proceedings of 6th Conference on Applied Natural Language Processing*, Seattle, pp. 7-12, 2000.

[14] Hutchins J., "Towards A Definition of Example-Based Machine Translation," *in Proceedings of the 2nd Workshop on Example-Based Machine Translation at MT Summit X*, Phuket, pp. 63-70, 2005.

[15] Hutchins W., *in The Encyclopedia of Languages and Linguistics*, Pergamon Press, 1994.

[16] Oflazer K., Çetinoğlu Ö., and Say B., "Integrating Morphology with Multi-Word Expression Processing in Turkish," *in Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, Barcelona, pp. 64-71, 2004.

[17] Orhun M., Adali E., and Tantuğ A., "Uygurcadan Türkçeye bilgisayarlı çeviri," *ITU Journal Series D: Engineering*, vol. 10, no. 3, pp. 3-14, 2011.

[18] Papineni K., Roukos S., Ward T., Zhu W., and Heights Y., "IBM Research Report Bleu : a Method for Automatic Evaluation of Machine Translation," *in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, pp. 311-318, 2002.

[19] Salymjanov I., Washington J., and Tyers F., "A Free/Open-Source Kazakh-Tatar Machine Translation System," *in Proceedings of the XIV Machine Translation Summit*, Nice, pp. 175-182, 2013.

[20] Shylov M., "Turkish and Turkmen Morphological Analyzer and Machine Translation Program," Masters Thesis, Fatih University İstanbul Turkey, 2008.

[21] SIL International, "Ethnologue: Languages of the World," [Online]. Available: http://www.ethnologue.com/family/17-15, Last Visited, 2016.

[22] Tantuğ A., "Akraba Ve Bitişken Diller Arasında Bilgisayarlı Çeviri Için Karma Bir Model," Thesis, Istanbul Technical University, 2007.

[23] Tantuğ A., Adali E., and Oflazer K., "Türkmenceden Türkçeye Bilgisayarlı Metin Çevirisi," *İstanbul Üniversitesi Mühendislik Derg*, vol. 7, no. 4, pp. 83-94, 2008.

[24] Tantuğ A., Adalı E., and Oflazer K., "A MT System from Turkmen to Turkish Employing Finite State and Statistical Methods Turkish and Turkmen Languages," *in Proceedings of MT Summit XI*, no. 1993, pp. 459-465, 2007.

[25] Tantuğ A., Oflazer K., and El-Kahlout I., "BLEU+: A Tool for Fine-Grained BLEU Computation," *in Proceedings of the International Conference on Language Resources and Evaluation*, Marrakech, pp. 1493-1499, 2008.

[26] Tayirova N., Tekerek M., and Brimkulov U., "Kırgız ve Türkiye Türkçeleri Arasında Istatistiksel Bilgisayarlı Çeviri Uygulaması Ve Başarım Testi," *MANAS Journal of Engineering*, vol. 3, no. 2, pp. 59-68, 2015.

[27] Venkatapathy S. and Joshi A., "Using

Information about Multi-Word Expressions for the Word-Alignment Task," *in Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, pp. 20-27, 2006.

[28] Wehrli E., Nerima L., and Scherrer Y., "Deep Linguistic Multilingual Translation and Bilingual Dictionaries," *in Proceedings of the 4th Workshop on Statistical Machine Translation*, Athens, pp. 90-94, 2009.

[29] Zhang Y., Vogel S., and Waibel A., "Interpreting BLEU/NIST scores: How much Improvement Do We Need to Have A Better System," *in Proceedings of Language Resources and Evaluation*, Lisbon, pp. 2051-2054, 2004.

Emel Alkim received her B.Sc., M.Sc. and Ph.D. in Computer Engineering from Dokuz Eylul University, Izmir, Turkey. Her main research areas are natural language processing and machine translation.

Yalçın Çebi received his B.Sc., M.Sc. and Ph.D. in Mining Engineering from Dokuz Eylul University, Izmir, Turkey. His main research areas include natural language processing, machine translation and wireless sensor and actor networks.