1. INTRODUCTION

This paper presents a rule-based machine translation (RBMT) system for the Kurdish language. The goals of this paper are two-fold: First, we build MT system using a free/open-source platform (Apertium). Second, we evaluate the translation of proposed system with “inkurdish” translation for the same set of data through manual evaluation method.

The Kurdish language belongs to the group of Indo-European languages. The Kurdish dialects are divided, according to the linguistic and geographical facts, into four main dialects. They are the North Kurmanji, Middle Kurmanji, South Kurmanji, and Gurani [1]. Kurdish is written using four different scripts, which are modified Persian/Arabic, Latin, Yekgirtu (unified), and Cyrillic. Latin script uses a single character while Persian/Arabic and Yekgirtu in a few cases use two characters for one letter. The Persian/Arabic script is even more complex with its RTL and concatenated writing style [2].

MT, perhaps the earliest NLP application, is the translation of text units from one natural language to another using computers [3]. Achieving error-free translation is a difficult task, instead an improvement in completely automatic, high quality, and general-purpose translations is required. The better MT evaluation metrics will be surely helpful to the development of better MT systems [4]. The MT evaluation has both automatic and manual (human) evaluation methods; the human evaluation criteria include the fluency, adequacy, intelligibility, fidelity, informativeness, task-oriented measures, and post-editing. The automatic evaluation method criteria include precision, recall, F-measure, edit distance, word order, part of speech tag, sentence structures, phrase types, named entity, synonyms, paraphrase, semantic roles, and language models. For this work, manual evaluation method has been used to evaluate the accuracy of both the systems.
We have used a platform called Apertium; Apertium defines itself as a free/open-source MT platform, initially aimed at related-language pairs but expanded to deal with more divergent language pairs and provide a language-independent MT engine and tools to manage the linguistic data [5].

Apertium originated as one of the MT engines in the project OpenTrad, which was funded by the Spanish government and developed by the Transducens research group at the Universitat d’ Alacanat. At present, Apertium has released 40 stable language pairs. Being an open-source project, Apertium provides tools for potential developers to build their own language pair and contribute to the project. Although Translators without Borders (TWB) claimed that they have developed offline MT engines for Sorani and Kurmanji, specifically for translating content for refugees using apertium, their work had not been published academically. Although Apertium was founded initially to provide an English/Catalan converter, it can also be used to right to left languages with more efforts specifically in creating transfer rules.

The rest of this paper is organized in the following way: Next, we present MT survey in Section 2. We describe methodology in Section 3. We then show and explain the results in Section 4, followed by the conclusion in the last section.

2. MT SURVEY

2.1. General MT Survey

A very early MT system returned to 1950s [6]. The development of computer with high storage and performance in one side and availability of bilingual and multilingual corpora in other side led to gain rapid MT development since the 1990s [7]. In 1993, IBM Watson research group did many important achievements in MT areas such as designing five statistical MT models and the techniques to estimate the model parameters using bilingual corpora [8]. In 2003, Franz Josef presented minimum error rate training for statistical MT systems [9] and Koehn et al., proposed statistical RBMT model [10]; in 2005, Koehn and Monz presented a shared task of building statistical MT systems for four European languages [11], and David Chiang proposed a hierarchical phrase-based SMT model that is learned from a bitext without syntactic information [12]; Menezes et al. used global reordering and dependency tree to build English-to-Spanish statistical MT in 2006 [13]. In 2007, Koehn et al. did a great achievement which was developing Moses, an open-source SMT software toolkit [14]; at the same time, in the sake of improving word alignment and language model quality among different languages, Hwang et al. team utilized the shallow linguistic knowledge [15]; Sa’lichez-Martínez and Forcada described an unsupervised method for the automatic inference of structural transfer rules for a shallow-transfer MT system in 2009 [16]. In 2011, Khalilov and Fonollosa designed a new syntax-based reordering technique to determine the problem of word ordering [17].

Deep learning fast development played a great roles in MT research evolving from conventional models to example-based models by Nirenburg in 1989 [18], statistical models by Carl and Way in 2005 [19], hybrid models by Koehn and Knight in 2009 [20], and recent years’ Neural models by Bahdanau et al. in 2014 [21].

Neural MT (NMT) is a recently hot topic that leads the automatic translation to be worked in a very different direction with the traditional phrase-based SMT methods. In traditional model, the different MT components are training separately, while the NMT components are training jointly by utilizing artificial neural network to increase the translation performance through two step recurrent neural network of encoder and decoder [21]-[23].

2.2. Kurdish MT Survey

Unfortunately, few efforts have been done for Kurdish MT yet. In 2011, Safeen Ghafour proposed a project called Speculate; Speculate can be considered as a theoretical research, a multiuse translator [24]. In 2013, the first English to Kurdish (sorani) MT system has been released under the name “inkurdish” for translating English text to Kurdish language [25]. In 2016, Google translate has added support for 13 new languages including Kurdish (Kurmanji dialect) language, bringing the total number of supported tongues to 10 [26]. TWB has developed offline MT engines for Sorani and Kurmanji, specifically for translating content for refugees [27]; in 2017, Kanaan and Fatima have evaluated “inkurdish” MT system using different automatic evaluation metrics in the sake of touching the weaknesses of “inkurdish” MT system [28]; Hassani suggested a method for MT among two Kurdish dialects (Kurmanji and Sorani) using bidialectal dictionaries, and his result showed that the translated texts are in 71% and 79% of cases rated as understandable for Kurmanji and Sorani, respectively. They are rated as slightly understandable in 29% of cases for Kurmanji and 21% for Sorani [2].
3. METHODOLOGY

The nature of language and availability of resources play important roles in selecting MT approach. Fig. 1 describes the four different categories of machine translation approaches.

### 3.1. Direct Translation

Direct translation involves a word-by-word translation approach. No intermediate representation is produced.

### 3.2. Rule-based Translation

RBMT systems parse the source text and produce an intermediate representation. The target language text is generated from the intermediate representation.

### 3.3. Corpus-based Translation

The advantages of this approach are that they are fully automatic and require less human labor. However, they require sentence-aligned parallel text for each language pair and cannot be used for language pairs, for which such corpora do not exist.

### 3.4. Knowledge-based Translation

This kind of system is centered around “Concept” lexicon representation a domain.

Rule-based approach has been chosen for this proposed system; reasons to choose a rule-based instead of a statistic system depend on the unavailability of sufficiently large corpora [29]; we use a RBMT which is suitable for languages, for which there are very little data [27]; despite being spoken by about 30 million people in different countries, Kurdish is among less-resourced languages [2]. Hence, RBMT is a suitable choice for Kurdish MT. RBMT models transform the input structure to produce a representation which matches the target language rules, and it has three components (Fig 2):

- Analysis, to produce the structure of source language;
- transfer, to transfer the representation of source language to representation of a target language; and generation, using target level structure to generate target language text.

After completing the prototype of the system, 500 different random data sets (simple sentence, complex sentence, proverbs, idioms, and phrases) have been given to both systems. Then, the output of both systems has been given to an annotator (English specialist - Kurdish native), to evaluate the results through manual evaluation method. The aim of the evaluation is to determine the translation accuracy for both systems in both meaning and grammar correctness. The evaluation has been designed into 5 categories, from score 5–1: Highly accurate, the translation is very near to the reference, it conveys the content of the input sentence, and no post editing is required; accurate, the translation conveys the content of the input sentence, and little post-editing...
is required; fairly accurate, while the translation generally
conveys the meaning of the input sentence, it suffers from
word order problems or tense or un-translated words; poorly
accurate, while the translation somehow conveys the meaning
of the input sentence, it does not convey the input sentence
content accurately; and completely inaccurate, the content of
the input sentence is not conveyed at all by the translation,
and it just give the translation of the words individually.

4. PROPOSED SYSTEM CONFIGURATION

Our system basically works on dictionaries and transfer rules,
and at a basic level, we maintain three main dictionaries:

1. Kurdish morphological dictionary: This file describes the
   rules of how words in Kurdish language are inflected,
   and its named: Apertium-kur.kur.dix
2. English morphological dictionary: This file describes the
   rules of how words in English language are inflected,
   and its named: Apertium-eng.eng.dix
3. Bilingual Dictionary: This file describes correspondences
   between words and symbols in Kurdish and English
   languages, and its named: Apertium-kur-eng.kur-eng.dix.

We maintain files for transfer rules in the two languages.
The rules govern the words reordering in target language,
the file is:

- English to Kurdish language transfer rules: This file contains
  rules govern how English will be changed into Kurdish
  language, its named: Apertium-eng-kur.kur-eng.t1x.

In spite of the possibility of translating Kurdish to English
texts, we just present English to Kurdish translation in this
work.

4.1. Terms Used in the System

Before creating the dictionaries and rules, some related terms
would be explained briefly. The first term is lemma: Lemma
is the form of word which is stripped of any grammatical
information, for example book is the lemma of (booked,
booking, etc.) and be is the lemma of was. The second term
is symbol: A grammatical label for example singular and plural
names, first person and present indicative, etc. Tags are used for
symbols, <n> for noun, <pl> for plural, etc. paradigm
is the another related term which refers to inflection of a
particular group of words, for example happy, happ (y, ier,
iest), instead of storing a lot of the same thing, We can simply
store one, and then we say the second inflects like the first,
for example “shy, inflects like happy”. Paradigms are defined
in <pardef> tags, and used in <par> tags.

4.2. Basic Tags in Kurdish and English Dictionaries

<Dictionary><dictionary/> tag is the start and end
point which contains the other all tags within xml file.
<Alphabet><alphabet/> tag defines the set of letters that
will be used in the dictionary.

<alphabet>ABCDEFGHJKLMNOPQRSTUVWXYZ
Zabcdefghijklmnopqrstuvwxyz<alphabet/> for English
dictionary.

<alphabet>رد خ ح چ ج ت پ ب اـئ
ێ ی وو ۆ و ە ـه ن م ڵ ل گ ک ق ڤ ف غ ع ش س ژ ز<alphabet/>
for Kurdish dictionary.

Symbol definitions: The symbols name can be written out in
full or in abbreviate, for example, noun (n) in singular (sg)
and plural (pl) (Fig. 3).

Then, we define a section <section><section/> for the
paradigms <pardef><pardef/> (Fig. 4).

This is the basic skeleton for the morphological dictionaries,
then the words will be entered through the entries,
<e><p><l/><r><s n="n"/><s n="sg"/></r></>

Fig. 3. Tags used for symbols.

<? xml version="1.0" encoding="UTF-8?>
<dictionary>
<sdefs>
   <sdef n="/n"/>
   <sdef sg="/sg"/>
   <sdef pl="/pl"/>
   <pardefs/>
</dictionary>

<section id="main" type="standard">
</section/>
<dictionary/>

Fig. 4. Skeleton for morphological dictionary.
p</e>, here e for entry, p for pair, l for left, and r for right. Compiling entries left to right lead to produce analyses from words and from right to left leads to produces words from analyses. The final step is compiling and run the dictionary. Both English (apergium-eng-eng.dix) and Kurdish (apergium-kur.kur.dix) morphological dictionaries would be created in the same manner.

### 4.3. Bilingual Dictionary
This describes mappings between words, the basic skeleton is the same as monolingual dictionary, but we need to add an entry to translate between the two words:

<e><p><l>university</s n="n"/></l><r>وکنانز</r></p></e>. We compile the bilingual dictionary left to right to produce the Kurdish → English dictionary and right to left to produce the English → Kurdish dictionary.

### 4.4. Transfer Rules
It contains rules to govern how English will be changed into Kurdish language, and the basic skeleton of the transfer rules is shown here (Fig. 5).

<rule> tag defines a rule. <pattern> tag means: “Apply this rule, if this pattern is found” (Here the pattern consists of a single noun defined by the category item nom). Patterns are matched in a longest-match first. The pattern matched and rule executed would be the first one. For each pattern, there is an associated action, which produces an associated output, out. The output is a lexical unit (lu). The <clip> tag allows a user to select and manipulate attributes and parts of the source language (side="sl") or target language (side="tl") lexical item. Transfer rules file need to be compiled and tested.

### 5. RESULTS AND DISCUSSIONS
After completing the prototype of the proposed system, it would be tested against different sets of data; first, we test it against individual words, and then simple sentence, complex sentences, phrases, proverbs, and idioms, some examples are shown in Fig. 6.

Fig. 6 shows a random sample of data translated by our proposed system; we tried to maintain a rich corpus that involves vast numbers of individual words, and then simple sentence, complex sentences, phrases, proverbs, and idioms, some examples are shown in Fig. 6.

The second part of this work will be evaluation between the proposed system’s results with “inkurdish” MT system results for the same set of data using manual evaluation method. Table 1 shows a sample of data translated by both systems. Inkurdish non-sense output with paragraphs and long texts obliged us to be satisfied at basic level (simple and compound sentence, idioms, proverbs, and phrases) evaluation; the sample contains a couple of random examples of each data set. The evaluation made by a neutral annotator (Kurdish native which is English specialist) according to the five categories has been defined before.

Detailed explanation of both computational and linguistics issues is out of our main aim, and we focused on accuracy differences between both systems, plus touching some general translation issues found during experimenting the data sets. Inkurdish MT system suffers severely from some issues, it is unable to link verbs to objects in sentences, and in spite of having all different meaning for a specific verb
in the corpus, it failures to bring the correct meaning of the verb according to its position in the sentence; it translated the verb “play” in “He went to play football before 1 h” example (Table 1) as ‘تێنیبەد ڵۆڕ’ instead of ‘تاكەد ىراي’ and this led to improper translation; the corpus of inkurish suffers from lack of pre defined common English idioms and proverbs; it always gives literal translation for idioms and proverbs for example, it translated “Better late than never” proverb to ‘زیگرەه ەل گنەرد رتشاب’ (Table 1) which is very literal and non-sense translation. Untranslated word is another issue for inkurish system for example the word “backyard” has not been translated in “The kids are playing in the backyard” example (Table 1).

Table 2 shows the accuracy average for all different data sets of both systems, and the accuracy averages have been calculated through a simple formula: Average = summation of all individual scores/total number of samples. The results showed that our system is more accurate than inkurish system for all data sets; both systems got high scores with “simple sentence” translation (3.12 and 3.56 of 5 for inkurish and our system, respectively); inkurish got the least score for idioms while our system for phrases (1.15 and

The accuracy average for both systems is shown in Table 2. Our system is more accurate than inkurish system for all data sets; both systems got high scores with “simple sentence” translation (3.12 and 3.56 of 5 for inkurish and our system, respectively); inkurish got the least score for idioms while our system for phrases (1.15 and 2.13 of 5, respectively), this means that inkurish needs to maintain large number of common English proverbs and idioms with their Kurdish equivalents while our system need to involve more English phrases.

In our previous work “Evaluation of inkurish MT System,” we addressed the issues of this MT system in details; hence, we tried to bridge the gaps of inkurish system in our proposed system and this is the reason of clear differences between inkurish accuracy average and proposed system accuracy average; the most common inkurish issue is lack of rich corpus specifically to deal with phrases, idioms, and proverbs (1.46, 1.15, and 1.25, respectively) (Table 2); during
experimenting the data with inkurdish, it did not translate even one idiom or proverb, and it gives a literal translation instead.

6. CONCLUSION

MT remains to be one of the most challenging aspects of NLP. Despite the ongoing efforts to achieve full machine-based translation, little progress has been achieved; due to language structure and composition complexity. Open-source platforms have provided the environment and tools required to develop reliable MT systems, especially for language with poor resources such as Kurdish. We have presented a MT system to translate English to Kurdish developed using an open-source platform. The resulting translation is compared with the result generated by inkurdish popular English to Kurdish MT system. The result shows clear differences between inkurdish MT system and our MT system in terms of translation accuracy. The result also shows that RBMT and manual MT evaluation are suitable choices, for poorly resourced languages.

Biography
Kanaan M. Kaka-Khan is an associate professor in the Computer Science department at Human Development University, Sulaimaniya, Iraq. Born in Iraq 1982. Kanaan M. Khan had his bachelor degree in Computer Science from Sulaimaniya University and Master Degree in IT from BAM University, India. His research interest area include: Natural Language Processing, MT, Chatbot, and Information Security.

REFERENCES

