

# Removing the Gender and Tense Discrepancies in Online English to Urdu Translators

\*Tariq Naeem<sup>1</sup>, Shanza Salomi<sup>2</sup>, Aman Ullah Khan<sup>3</sup>

<sup>1,2,3</sup> Department of Computer Science & Engineering, Air University Multan Campus, Pakistan

<sup>1</sup>naeemtarik@aumc.edu.pk (\*corresponding author), <sup>2</sup>shanza.atta@outlook.com, <sup>3</sup>aak@aumc.edu.pk

## Abstract

Existing translators like Google Translate, Bing Microsoft Translator and Collins Translator does not identify gender and tense cases in translation from English to Urdu. This paper, primarily based on, ruled based methodology to machine translation for English (source language) to Urdu (target language), handling tense identification and semantic translation of gender cases. Data was collected gathered from published resources like BBC Urdu, newspapers, magazines, novels and literature. POS (Part of Speech) tag model and POS-based reordering techniques are presented. Analysis and findings segment of the research article elucidated our testified outcomes in detail. The proposed approach achieved 79% accuracy. The developed application allows contributing towards information to comprehend writing, logical inquires and literature in Urdu language and translate it in equivalent English language.

Keywords: Online Translators, Rule based Approach, Gender and Tense Case

## 1. Introduction

Natural Language Processing (NLP) has been now introduced as an interdisciplinary course in the fields of artificial intelligence, linguistics and computer science that is very helpful in exploring the usage of computers in understanding and manipulating the text or speech of natural language [1]. Computer software automatically translates the text using Machine Translation (MT) methods [3]. In MT, the source natural language is converted automatically into a new-targeted language, however, keeping the meaning of the input text original and fluent in the output language, as shown in Fig 1.

There are three architectural classifications of MT systems i.e. Direct, Transfer and Interlingua architectural [8]. In addition, RBMT, EBMT, SMT and hybrid are the approaches of MT system [5].

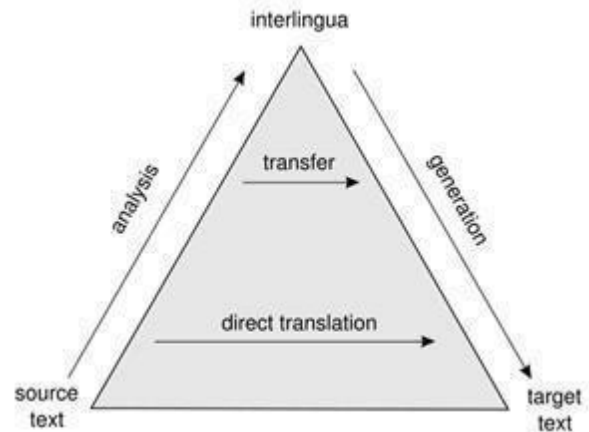


Fig 1: Machine Translation Architecture [8]

Direct Architecture is the basic type of translation substituting source language (SL) lexeme with the target language (TL) lexeme; Interlingua approach necessities a top to bottom semantic investigation and the TL is produced through this methodology. Transfer approach lies in the middle of the two boundaries; it deals with the syntactic dimension and includes semantics in a few spots. The syntactic arrangement of SL is explored to develop a processed structure, plot the rules to change it into TL arrangement, and interpretation is then produced with the use of TL definite rules.

MT provides many benefits like without the human translators a large amount of text is converted from one natural language to other language which reduce spending of money and time with less human efforts [7] Google Translate is an online translator that is a free of cost text translation system, which implies the

statistical machine translation patterns and provides translations for 55 different languages or more. Besides this Microsoft Translator is also available which implies EBMT and different SMT translation systems. RBMT system is implied by Systran. English, Chinese, Arabic, Dutch, French and other languages are converted by Systran. Most of the other languages worked in pairs with English or French [18]. Another one Babel Fish is a web based translator that supports multilingual translation. Babel Fish translator provides services to translate webpages among 36 pairs and 13 languages including, Russian, Spanish, Korean, English, Dutch, French, Japanese, Traditional Chinese, Italian, Simplified Chinese, German, Greek and Portuguese [2]. The ImTranslator is a free web-based translator that performs translation of words, text or phrases between more than 90 languages. ImTranslator uses statistical machine translation (SMT), this online translator uses statistical methods that is based on multilingual texts. Statistical machine translation uses the existing translation (done by human translators) of source and target language for making new rules to translate between languages. The accuracy of the translator is increased by using statistical machine translation approach.

Issues and problems do exist in machine translations that make it difficult such as order of words, idioms, ambiguity in word sense, preposition, post-position, awareness of gender and context because natural languages are very complex. There are multiple meanings of most words, various readings are available for sentences and grammatical rules of one language may differ from other languages. Besides this, there are other non-linguistic aspects such as the word knowledge and language morphology is needed for carrying out translation [6] and [7].

The rest of paper is composed as follows. Section 2 explains the literature review and related work. Section 3 illustrates the strategies and methodology. Results and discussion about current work is illustrated in section 4. Section 5 closes the article with future work indicators.

## 2. Literature Review

Development of framework for translation from English to Urdu is considerably insufficient in comparison with the number of Urdu speakers [13].

More work is highly needed in this field. Urdu has been ranked at 19<sup>th</sup> number out of 7,105 languages spoken all over the world. It is one of the most common languages in South Asian region [11]. About 5% to 10% people only can speak or understand English in Pakistan [9], [21] and [10].

According to [12], for developing and implementing a translator all grammar of SL must be developed with bottom-up parsing algorithms. After acquiring the parse tree of SL sentences, it is then translated according to those grammatical rules in to TL sentences. In [14], corpus-based MT system was proposed to tackle problems like syntactic and structural ambiguity, anaphoric resolution and discourse analysis using data mining and text mining tools.

MT system for English to Bodo [5] uses general domain English-Bodo parallel text corpora. But the computational system containing information of Bodo language was not enough and it required more expansion.

The authors of [15], proposed a method for translating Malayalam text into English. For this purpose, rule-based machine translation system was used. The system comprises of bilingual dictionaries and conversion rules. These rules were implied for text conversion from SL to TL. In case of multiple meanings in English of a Malayalam word the proposed system generated multiple sentences.

Whereas, for translating English text into Urdu an expert system using Unicode Standards for translation was proposed in [16]. The Unicode worked with a knowledge base which contained grammatical patterns of English and Urdu, as well as a tense and gender-aware dictionary of Urdu words (with their English equivalent forms). In order to avoid the problems occurred in case of multiple meaning of a single word AGHAZ solution was implemented. A parsing-based reordering technique was presented in [17] which were used for English-to-Japanese phrase translation. The phrase-based translator is used to increase the performance of translation through reordering technique. The reordering technique was also used in the preprocessing stage for syntax based translation.

In [18], the author focused on the rule-based case transfer, as shown in Fig 2, which was a part of the transfer grammar module developed for bidirectional Tamil to Malayalam MT system. The presented study

involved two typologically close and genetically related languages, Tamil and Malayalam. They considered the basic construction of sentences which was highly dependent on the case systems. The rules were written by taking into consideration the postpositions and cases in the languages. A parallel corpus was chosen and deep analyses of the case transfer patterns were drawn and rules were written to sort out the case changes that happen when translating from SL to TL. Web data was used for evaluation and the results were encouraging.

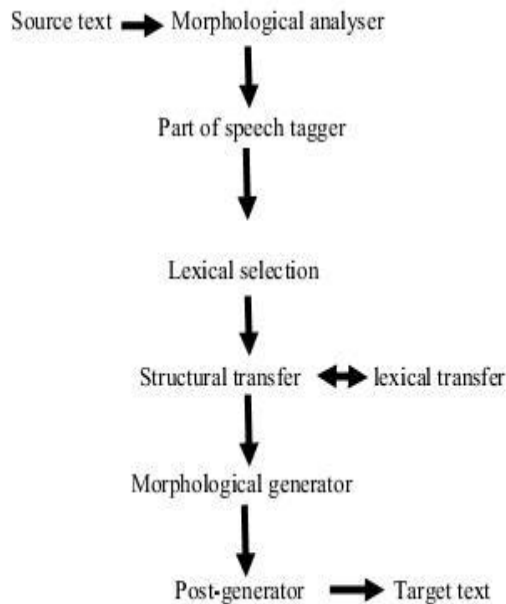


Fig 2: Architecture of RBMT [18]

As [19] presented, a system outline of English to Hindi Machine-Aided translation system named AnglaHindi. This translation system utilizing rule-based and example-based approaches, with some statistics to achieve more satisfactory and precise translation for regular verbs and nouns phrases. This approach to some degree combined hybridization of rule-based and example-based.

A translator that translate English text into Arabic by using rule based machine translation approaches was also used feed-forward back propagation of Artificial Neural Network (ANN) was proposed by [4], [20] and [25]. The proposed system translates sentences that have prepositional objects, gerunds,

direct and indirect objects, infinitives, etc. Neural networks worked with bilingual dictionaries that do not have the meanings of English words into Arabic but it also store all the linguistic details of words of one language to other languages.

A translator proposed by [9] to translate English text into Urdu text through the use of example based MT. The English and Urdu translator supported homographs, idioms, and helped out other features that had the aptitude of the bilingual corpus to grow.

Some examples from traditional MT systems are discussed below. Several sentences were taken from books, magazines, and online forum to test the gender and tense cases.

**She is playing.**

(1)

- [22] وہ کھیل رہا ہے۔
- [23] وہ چلا رہی ہے۔
- [24] وہ چلا رہی ہے۔

**Amna is my friend.**

(2)

- [22] امینہ میرے دوست ہے۔
- [23] آمنہ میرا دوست ہے۔
- [24] آمنہ میرا دوست ہے۔

**The boys wanted to help him.**

(3)

- [22] لڑکوں کو اس کی مدد کرنا چاہتا تھا۔
- [23] لڑکوں نے اس کی مدد کرنا چاہتا تھا۔
- [24] لڑکوں نے اس کی مدد کرنا چاہتا تھا۔

In these above statements, Google Translator [22], Microsoft Bing Translator [23] and Collins Translator [24] showed semantically incorrect translation output. A huge demand comes from users who are not familiar with English to build an automated system, which can cope with such issues. For this, the English to Urdu machine translator is designed and developed, which is a web-based system for providing correct language translation semantically.

### 3. Methodology

Natural Language Processing Toolkit (NLTK) is a main platform for structuring Python programs to work with language data. It gives simple interfaces to more than fifty corpora and lexical resources such as Word Net. It also provides a suite of content handling libraries for stemming, tagging, tokenization, classification and semantic reasoning.

Apart from having a discussion forum, NLTK also cover the highly develop industrial libraries using recent NLP methodologies.

```
from textblob import TextBlob
from nltk import word_tokenize, pos_tag
from pip._vendor.distlib.compat import raw_input
engtext=raw_input("Please type your english sentence:")
print (engtext)
blobObj=TextBlob(engtext)
print(blobObj.tags)
```

Fig 3: Code for English POS tags

TextBlob is a python library utilizing NLTK. Practically, all required tasks needed in essential NLP works well as a framework with TextBlob.

Apart from it, TextBlob has some advance features. For instance, Part-of-speech tagging (see Fig 3), sentiment analysis, noun phrase extraction, classification (Naive Bayes, Decision Tree), lemmatization, language transis a leading platlation by Google, word and phrase frequencies, tokenization (splitting text into words and sentences), word inflection (singularization and pluralization) and spelling correction.

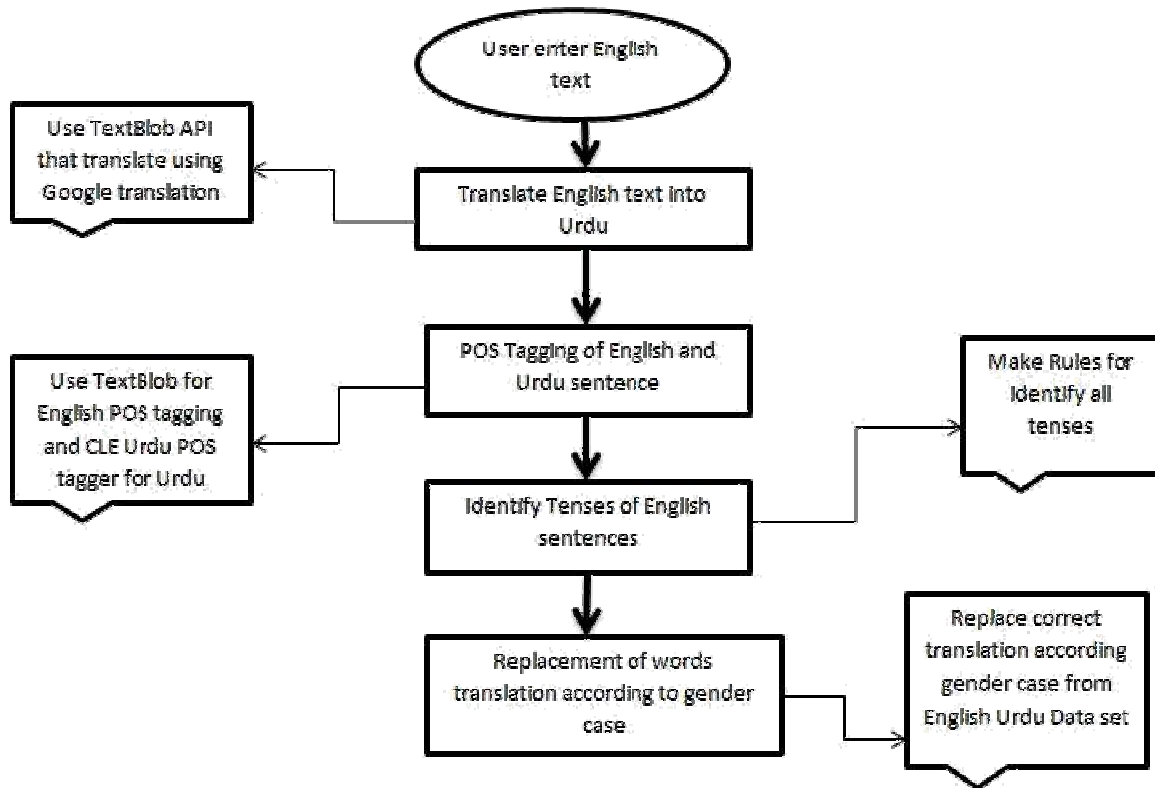
```
import json
import requests
from nltk import word_tokenize, pos_tag
from pip._vendor.distlib.compat import raw_input
from textblob import TextBlob
engtext=raw_input("Please type your english sentence:")
print (engtext)
blobObj=TextBlob(engtext)
z=blobObj.translate(from_lang="en", to="ur")
print("translation english to urdu:", z)
accessToken = "583e203b-11ed-42bc-893f-ff8459fc6d56"
inputText= str(z)
data = json.dumps({'text': inputText, 'token': accessToken})
headers = {'content-type': 'application/json'}
resp = requests.post('https://api.cle.org.pk/v1/pos', data=data,
headers=headers)
print(resp.json())
```

Fig 4: Code for Urdu POS tags

TextBlob translate any English sentences in Urdu that is entered by user see Fig 4. TextBlob have the feature of POS (part of speech) tagging on English sentence. For instance:

**She goes for a walk daily.** (4)  
[('She', 'PRP'), ('goes', 'VBZ'), ('for', 'IN'), ('a', 'DT'), ('walk', 'JJ'), ('daily', 'NN')]  
{'response': {'status': 'ok', 'tagged\_text': 'PRP| وہ RB| روزانہ VBF| چلتا NN| بے '}}

**The boys wanted to help him.** (5)  
[('The', 'DT'), ('boys', 'NNS'), ('wanted', 'VBD'), ('to', 'TO'), ('help', 'VB'), ('him', 'PRP')] {'response': {'status': 'ok', 'tagged\_text': 'NN| لڑکوں PSP کو PRP| اس PSP| کی NN| تھا NN| چاہتا AUXM| کرنا VBI| مدد NN}}



**Fig 5:** Implementation of English to Urdu Translator

In these sentences, TextBlob Part of Speech tagging technique separates the POS tags. After POS tagging for Urdu sentences were applied. The proposed system, shown in Fig 5, use Urdu POS tag set of CLE (Center for Language Engineering) for Urdu part of speech tagging [26]. Urdu tag set of CLE

contributions as “Urdu word sense annotation tool” is developed to run a simple interface for word sense labeling and confirming labeling stability. After identifying of tenses, the data set for English Urdu meanings was developed then the incorrect translation with correct translation was replaced.

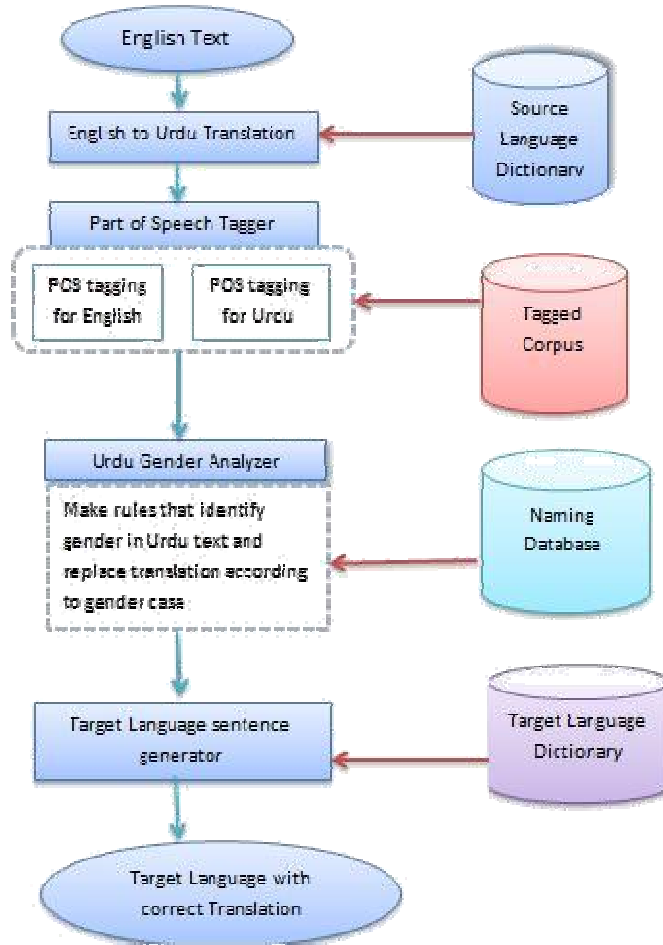


Fig 6a: System Framework for Gender Analysis

We have presented our framework in which we have explained a process of part of speech tagging of English text. Identification of gender and tense case of a given English text is shown. A clear view of the tense identification in framework that is illustrated above in Figure 6.

#### 4. Result and Discussion

We checked our English to Urdu translator with various sentences of English and Urdu language. From our proposed framework, we have corrected both tense and gender cases.

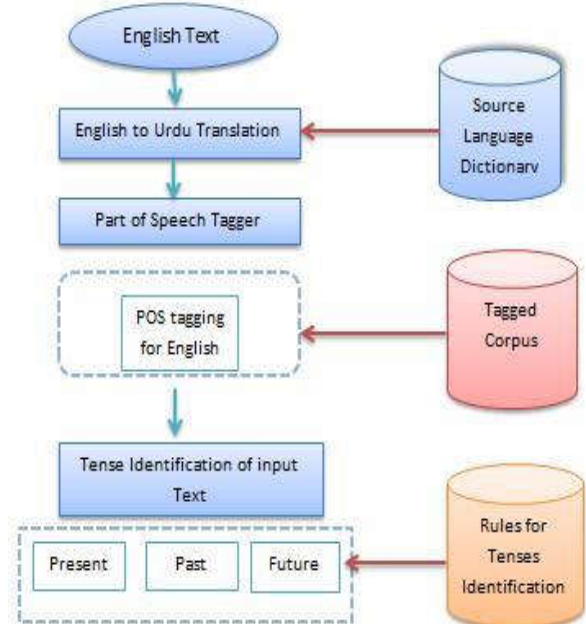


Fig 6b: System Framework for Tense Analysis

#### 4.1 Tense Case

In tense case we have tested our system with present, past and future tense sentences. They were accurately identified by our proposed method.

I am playing.

(6)  
میں کھیل رہا ہوں۔

```

C:\Users\Hp\PycharmProjects\DEMO\venv\Scripts\python.exe
Please type your english sentence:i am playing
i am playing
English: i am playing
میں کھیل رہا ہوں
[{'i', 'NN'}, {'am', 'VBP'}, {'playing', 'VBG'}]
{'future': 0, 'present': 2, 'past': 0}

Process finished with exit code 0
  
```

Fig 7: Present tense case by proposed system

**She worked in office for three years. (7)**

انہوں نے دفتر میں تین سال تک کام کیا۔

**We shall play together. (8)**

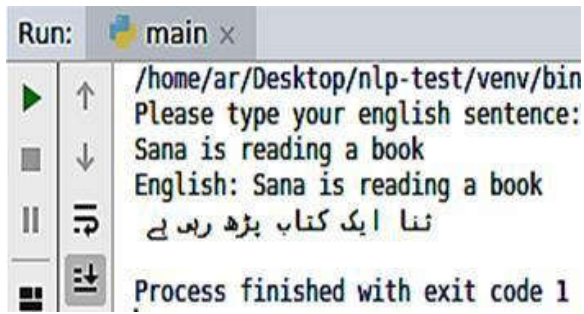
ہم ساتھ مل کر کھیلے گئے۔

## 4.2 Gender Case

Some examples of English sentences, whose incorrect translation by [22], [23] and [24] were corrected by our proposed system i.e.

**Sana is reading a book. (10)**

سنا ایک کتاب پڑھ رہی ہے



```
Run: main x
/home/ar/Desktop/nlp-test/venv/bin
Please type your english sentence:
Sana is reading a book
English: Sana is reading a book
سنا ایک کتاب پڑھ رہی ہے
Process finished with exit code 1
```

Fig 8: Gender case by proposed system

## 5. Error Analysis

A sample of 190 English sentences was taken from books, magazines and online platforms given as input to the proposed system. Out of 190 sentences, the system achieved semantically accurate translation of 150 sentences.

The accuracy of the system was calculated simply using the percentage formula, i.e.

$$(150/190) * 100 = 78.9\%$$

Due to long and complex English sentence structure, the system could not generate semantically correct translation.

## 6. Conclusion

This system remove the gender discrepancies in online English to Urdu translators because the existing translators like Google Translate, Bing Microsoft Translator and Collins Translator does not identify the gender (male or female) of the person names in

translation from English to Urdu. If we write the name of female in English text, Google translator gives the translation according to male gender but this research, primarily based on, ruled based approach of source language to target language, handling semantic translation of gender cases and tense identification. Our MT system supports POS (Part of Speech) tags models use for tagging of English and Urdu text and our English to

Urdu MT system gives accurate translation according to gender (Male or Female). The proposed system achieved 79% accuracy.

Although this not the first study in English to Urdu MT, however, less efforts are done to consider semantically gender cases during translation. The proposed system provide translation of simple English sentences into Urdu, we use rule based machine translation approach. The main objective was to identify these cases in English to Urdu MT system that never discussed before.

## 7. Future Work

The future work and extension of this work can be the extraction of accurate translation for complex long sentences. Because in long complex sentence structure a simple POS matching of one word may not work; especially if there are multiple verbs (different forms) in a sentence and one of them is incorrect. How would the authors know which one is wrong?

## 8. Reference

- [1]. Liu, D., Y. Li, and M.A. Thomas. *A roadmap for natural language processing research in information systems*. in *Proceedings of the 50th Hawaii International Conference on System Sciences*. 2017.
- [2]. Liddy, E., *Natural Language Processing, 2nd edn. Encyclopedia of Library and Information Science*. Marcel Decker. Inc., NY, 2001.
- [3]. Khan, S. and R. Mishra, *Translation rules and ANN based model for English to Urdu machine translation*. INFOCOMP, 2011. 10(3): p. 36-47.
- [4]. Antony, P., *Machine translation approaches and survey for Indian languages*. International Journal of Computational Linguistics & Chinese Language Processing, Volume 18, Number 1, March 2013, 2013. 18(1)

- [5]. Islam, S. and B. S. Purkayastha, *Implementation of English to Bodo Machine Translation System using SMT Approach*. International Journal of Computer Science & Applications, 2017. 14(2)
- [6]. Costa-Jussa, M.R., et al., Study and comparison of rule-based and statistical catalan-spanish machine translation systems. Computing and informatics, 2012. 31(2): p. 245-270
- [7]. Naeem T., Khan MA, Automatic derivation of Nouns from Adjectives, 6th International Conference on Language and Technology, 41-49, 2016
- [8]. Islam, S., M. Devi, and B. Purkayastha, *A study on various applications of NLP developed for North-East languages*. International Journal on Computer Science and Engineering, 2017. 9(6): p. 386-378.
- [9]. Alqudsi, A., N. Omar, and K. Shaker, *Arabic machine translation: a survey*. Artificial Intelligence Review, 2014. 42(4): p. 549-572.
- [10]. Zafar, M. and A. Masood, *Interactive english to urdu machine translation using example-based approach*. International Journal on Computer Science and Engineering, 2009. 1(3): p. 275-282.
- [11]. Hussain, S. *Urdu localization project: Lexicon, MT and TTS (ULP)*. in *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*. 2004, Association for Computational Linguistics
- [12]. Malik, A.A. and A. Habib. *Qualitative Analysis of Contemporary Urdu Machine Translation Systems*. in *NLPAR@LPNMR*. 2013. Citeseer.
- [13]. Ahmed, T. and S. Alvi, *English to Urdu translation system*. manuscript, University of Karachi, 2002.
- [14]. Revanuru, K., K. Turlapaty, and S. Rao, *Neural Machine Translation of Indian Languages*. 2017.
- [15]. Tahir R., Asghar S, and Masood. *Knowledge based machine translation*. in *Information and Emerging Technologies (ICIET), 2010 International Conference on*. 2010. IEEE.
- [16]. Rajan, R., et al. Rule based machine translation from english to malayalam. *Advances in Computing, Control, & Telecommunication Technologies*, 2009. IEEE.
- [17]. Muhammad, U., et al., Aghaz: An expert system based approach for the translation of english to urdu. *International Journal of Social Sciences*, 2008. 3(1): p. 70-74.
- [18]. Lee, Y.-S., B. Zhao, and X. Luo. Constituent reordering and syntax models for English-to-Japanese statistical machine translation. in *Proceedings of the 23rd international conference on computational linguistics*. 2010. Association for Computational Linguistics.
- [19]. Lakshmi, S. and S.L. Devi, Rule Based Case Transfer in Tamil-Malayalam Machine Translation. *Research in Computing Science*, 2014. 84: p. 41-52.
- [20]. Sinha, R. and A. Jain, Angla Due to long and complex English sentence structure, the system could not generate semantically correct translation. Hindi: an English to Hindi machine- aided translation system. *MT Summit IX, New Orleans, USA, 2003*: p. 494-497.
- [21]. Akeel, M. and R. Mishra, ANN and rule based method for english to arabic machine translation. *Int. Arab J. Inf. Technol.*, 2014. 11(4): p. 396-405.
- [22]. Google Translator, 2018, [online]. Available: <https://translate.google.com/>, [Accessed: 16-Dec-2018]
- [23]. Microsoft Bing Translator, 2018, [online]. Available: <https://www.bing.com/translator>, [Accessed: 16-Dec-2018]
- [24]. Collins Translator, 2018, [online]. Available: <https://www.collinsdictionary.com/translator>, [Accessed: 16-Dec-2018]
- [25]. Mohsin A., Asghar S., Naeem T., Intelligent Security Cycle: A rule based run time malicious code detection technique for SOAP messages, 19th IEEE International Multi-Topic Conference (INMIC), 1-10, 2016
- [26]. Urooj S., Shams S. Hussain S. Adeeba F.; Sense Tagged CLE Urdu Digest Corpus, CLE KICS, UET, 1-8, 2018