A Brief Study of Challenges in Machine Translation

H.Mohamed Zakir[1] and M.Shafeen Nagoor[2]

[1] Preparatory Year Deanship, King Faisal University, Hofuf,Alahsa,SaudiArabia

[2] Preparatory Year Deanship, King Faisal University, Hofuf,Alahsa,SaudiArabia

Abstract

Of late English has become one of the most preferred language worlds over. However, not everyone on this globe is a conversant with this medium of communication. Machine Translation has become indispensable in such a scenario where physical and logical boundaries are vanishing and one need to be able to communicate at will and in a medium which he is conversant with. At present Machine Translation is the most fascinating but equally a challenging problem. Researchers are trying to translate English language to their native language, but achieving a flawless Machine Translation has become a real challenge for researchers all over the world. This paper discusses various open challenges in machine translation with a focus on the problems encountered in English to Urdu Machine Translation. We also discuss the parallel corpora, which we feel is a key concept in Machine Translation and may provide a better solution to these open challenges in Machine Translation.

Keywords: *Machine Translation Challenges, Parallel Corpora.*

1. Introduction

Machine translation (MT) is automated translation of text by a Computer without any human participation. It is the process, by which computer programs are used to translate a text or sentence from one natural language (such as Urdu) to another natural language (such as English).

Apparently the first suggestions concerning Machine Translations (MT) were made by the Russian Smirnov-Troyansky and the French man G.B Artsouni in the 1930's. However the first serious discussions were begun in 1946 by the mathematician Warren Weaver. He and many others were inspired by the success of the allied efforts using the British Colossus computer to break the German military code produced by the Enigma machine, and the obvious similarity between the task of decoding and encoded message and the task of translation of one language into another. By 1954, there was a Machine Translation project at Georgetown University, which succeeded in correctly translating several sentences from Russian into English. Soon there were Machine Translation projects at MIT, Harvard and the University of Pennsylvania. [Thomas D. Hedden].

In 1964, after more than $20,000,000 had been invested by the Federal Government in MT, the National Academy of Sciences commissioned the Automatic Language Processing Advisory Committee (ALPAC) to write a study of the status of MT. The committee, headed by John R. Pierce, wrote a now-famous report in which it expressed doubt that a fully-automatic MT system could ever be produced. That report sounded the death-knell for funding of MT research, and MT was neglected for many years afterwards. [Thomas D. Hedden].

The reasons for this failure have been described many times, and come down to the fact that the analysis of messages by humans in natural language relies to some extent on information which is not present in the words which make up the message. This led the linguist Yehoshua Bar-Hillel to declare that MT was impossible. The example which he provided has since become a classic, and is now called the Bar-Hillel paradox:

The pen is in the box.
[I.e. the writing instrument is in the container]
The box is in the pen.
[I.e. the container is in the playpen or the pigpen]

There are two possible ways that a person could correctly infer the meaning of these sentences. First, if there is a context preceding these sentences, it could make clear which meaning of pen is being used in which sentence. That is, the meaning of the words and information about the context is carried over from one sentence to the next. There is now an entire branch of linguistics, called discourse analysis, devoted to the study of how context affects the meaning of words and sentences. In order to infer in this way the correct meaning of an ambiguous sentence, computers will have to learn how to "remember" a context and make use of it to interpret the correct meaning of words and sentences within that context.

However, in the examples given above, most humans can understand the meaning correctly without any context. In order for a fully automatic MT system to translate these sentences correctly, the following information would have to be available to the computer.

- Pens [writing instruments] are smaller than boxes.
- Boxes are bigger than pens [writing instruments], but smaller than pens [playpens, pigpens, etc.]
- It is impossible for a bigger object to be inside a smaller object

Thus, one way or the other, whether the correct meaning of the sentences is inferred based on the context or in isolation, it is necessary for the computer to have information at its disposal which is not included in the message itself. During the early days of MT this realization was enough to make MT seem an impossible task.

Interest in MT revived in the 1980's, following dramatic advances in computer hardware (storage capacity, speed, etc.) and software (LISP, etc.). The need to store and process tremendous amounts of real-world knowledge in order to analyze a single word in the message ceased to be an impediment to design and use of MT systems [Thomas D. Hedden].

Precise Machine Translation services such as Google Translate, Bing Translator etc. is the demand of time as it is required for communication, information sharing and for some other purposes, but unfortunately the computer scientists have yet to achieve this goal. Machine Translation poses certain challenges since exact translation of one language into another makes up a complicated problem. An attempt has been made to identify these Machine Translation challenges in English to Urdu Machine Translation and discussed below.

2. Challenges in Machine Translation

Researchers all over the world are looking for some permanent solutions for Machine Translation issues. On and off many Machine Translation challenges have been identified and have been demanding addressable. It needs a lot of concern and requires a keen observation to recognize and to resolve these Machine Translation problems. These problems were analyzed and categorized as under:

1. Word Translation Problems.
2. Phrase Translation Problems.
3. Syntactic Translation Problems.
4. Semantic Translation Problems.

2.1 Word Translation Problems

Proper word translation is one of the major challenges in Machine Translation. In many languages, a single word has multiple meanings, the same is the case in English and Urdu languages, so to find out the right meaning and thus the machine translation of an English word (with multiple meanings) into Urdu is a real challenge. The humans can understand the right meaning of a word (with multiple meanings) by looking at the context, but the machines (computers) are still unable to fix it up. Some of the examples are taken up from Google Translate which depicts the word translation problem.

Please book my ticket for tomorrow.

کل میرا ٹکٹ کتاب کری

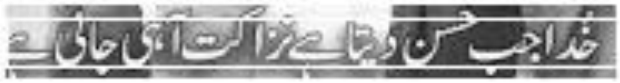
Please buy that book for me.

میرے لئے اس کتاب خریدنے کریں

In the above mentioned example, the word book has different meanings in two different sentences, in the first sentence, the word book means to reserve a seat in advance and in the second sentence the same word book means a written work or composition that has been published (printed on pages bounded together). But, Google Translate translates the word book in both the sentences as the later one (the published work).Many other English words with multiple meanings have been examined but all of them have the same translating issue. So the machine translation of words with multiple meanings is a real challenge and need to be addressed.

2.2 Phrase Translation Problems

Phrase translation is the challenge in English to Urdu machine translation. Idiomatic phrases have hidden meaning. We cannot translate the phrases word by word. So to get an appropriate Urdu translation of an English phrase is a big challenge in Machine Translation. We made an attempt to translate an English phrase ***"Beauty requires no ornaments***" into Urdu, using Google Translator and the result displayed on the screen by translator was absolutely wrong. This phrase must be translated as



خدا جب حسن دیتا ہے نزاکت آہی جاتی ہے

Fig.1 Expected Translation

But it was translated as,بیوٹی کوئی زیور کی ضرورت ہوتی ہے

By analyzing this example, we realized that phrase translation is still an existing challenge in Machine Translation. Research scholars are proposing different techniques and models for solving the phrase translation problem but the solution is yet to come. Hopefully one day very soon this Machine Translation problem will be resolved.

2.3 Syntactic Translation Problems

One more problem in English to Urdu machine translation is the syntactic translation problem. This problem occurs due to the various differences among the languages. It depends upon the degree of relatedness between the languages. Syntactical problems may decrease, if the languages belong to the same family. For Example, English and German belongs to the Indo-European family, Tamil and Telugu belongs to the Dravidian family.

English and Urdu, both languages use different syntax. English follows **Subject - Verb – Object** word order, however, Urdu has **Subject – Object – Verb** sentence structure. So this variation in the syntax of these languages (English and Urdu) leads to the syntactic translation problem. For further illustration of this type of problem an example is given below.

English: I bought a pen.

Subject Verb Object

Urdu: مجھے ایک قلم خریدا

Verb Object Subject

2.4 Semantic Translation Problems

One more translating challenge in English to Urdu Machine Translation is to resolve the pronominal anaphora problem (Pronoun Resolution). An anaphor is a word or phrase used to refer back to a previous word or phrase in the same text. For further illustration of Semantic Translation problem we have some examples given below:

1. My son dropped the glass plate and **it** broke into pieces (the glass plate)
2. The child wanted a toy but his father didn't buy **one** for him. (Toy)

In Machine Translation, while translating these sentences into Urdu we have to take care of pronoun resolution. The pronoun **it** could potentially refer to either the data or the computer, so how the machines should deal with this type of pronominal anaphora problem need to be resolved. The other related problem in machine translation includes coreference Translation problem and Discourse Translation problem.

In this paper, after discussing the challenges in English to Urdu machine translation we decided to include one more key concept in Machine Translation called Parallel Corpora, which we feel may provide a better solution to these translating challenges.

3. Parallel Corpus

A corpus is a large collection of texts, stored on a computer. A Parallel Corpus contains texts in two languages. Two main types of Parallel Corpus are given below.

Comparable corpus: the texts are of the same kind and cover the same content.

Translation corpus: the texts in one language (L1) are translations of texts in the other language (L2).

3.1Types of Parallel Corpora

Parallel corpora can be bilingual or multilingual, i.e. they consist of texts of two or more languages. They can be either unidirectional (e.g. an English text translated into German), bidirectional (e.g. an English text translated into German and vice versa), or multidirectional (e.g. an English text such as an EU regulation translated into German, Spanish, French, etc.).[Glottopedia]

3.2Compilation of Parallel Corpora

The texts of a corpus are chosen according to specific criteria which depend on the purpose for which it is created. In particular, compilers have to decide whether to include a static or dynamic collection of texts, and entire texts or text samples. Questions of authorship, size, topic, genre, medium and style have to be considered we well. In any case, a corpus is intended to comply with the following requirements: (i) it should contain authentic (naturally occurring) language data; (ii) it should be representative, i.e. it should contain data from different types of discourse. [Glottopedia]

3.3Alignment of a Parallel Corpus

In order to use a parallel corpus properly, it is necessary to align the source text and its translation(s). This means that one has to identify the pairs or sets of sentences, phrases and words in the original text and their correspondences in the other languages. Parallel text alignment is important because during the translation process sentences might be split, merged, deleted, inserted or reordered by the translator in order to create a natural translation in the target language. In order to compare the original text and its translation(s), it is necessary to (re-) establish the correspondences between the texts. In the process of alignment, anchor points such as proper names, numbers, quotation marks etc. are often used as a point of orientation. The degree of correspondence between the texts of a parallel corpus varies depending on the text type. For example, a fictional text may allow the translator a greater freedom than a legal one. [Glottopedia]

4. Future Work

After recognizing the existing translating problems in machine translation, we realized that this analysis will be helpful for our future research work. In future we have to focus on the design and development of a Machine Translation toolkit with special reference to tourism and hospitality management in our area. The main purpose for the development of the kit is the implementation of parallel corpora.

5. Conclusions

In this paper, we discussed the problems in English to Urdu Machine Translation with the help of some examples. We hope that our research work will definitely help those research scholars who are doing their research in Machine Translation. We also proposed the concept of parallel corpora which we feel may produce a better result to these open challenges in Machine Translation.

References

[1] Thomas D.Hedden, Machine Translation: A Brief Introduction, 1992-2010.
[2] Douglas Arnold, Lorna Balken, Siety Meijer, R. Lee Hunphrets, Louisa Saller. Machine Translation: An Introductory Guide.
[3] Tognini-Bonelli, 2001:70.
[4] Glottopedia, www.glottopedia.org/index.php/parallel_corpus

First Author The first author has around 9 years of experience in the field of Computer Science in teaching and the software industry. He holds a bachelor degree in Computer Science and Engineering (2006) and his Post graduation in Computer Science (2013).He started his career as a developer engineer in Wins Infotek Private Limited which is a Japan based company. Currently associated with King Faisal University as an IT Trainer. He published a journal titled “*Time Comparison Algorithm for University Examination Scheduling*” in International Journal of Scientific Research and Innovative Technology

Second Author The second author has around 7 years of experience in the field of Computer Science in teaching. She holds a bachelor degree in Computer Science and engineering (2006) and has dual Post graduation degrees in Business Administration (2010) and Computer Science (2014).Currently associated with King Faisal University as an IT Trainer.