CORPUS-BASED MACHINE TRANSLATION

Eassa Ali Mohammed Ali[1], Ameen Ali Mohammed Al-Gamal[2]
[1]PhD Research Scholar, Department of Linguistics, University of Kerala, English Language Instructor, Department of Translation, College of Languages, Sana'a University, Republic of Yemen
[2]PhD Research Scholar, Department of Linguistics, University of Kerala English Language Instructor, Department of English Language, Faculty of Education, Thamar University, Republic of Yemen
University of Kerala, Trivandrum, India

Received: May 25, 2021 Accepted: July 01, 2021

ABSTRACT: *A Corpus is a collection of texts either written or spoken, and it is stored for the purpose of analysis. A corpus can be used for multiple purposes such as language learning, machine translation, and etc. Corpus-based machine translation refers to a situation in which the software depends on a stored data 'corpus' to translate from one language into another language. The Corpus-based machine translators may be able to handle the linguistic issues, but it may be difficult to translate the cultural items and concepts. Therefore, they cannot replace human translators; rather they can only assist them. This study discusses the corpus-based machine translators compared to human translators with regards totranslating linguistic and cultural items from Arabic into English and vice versa.*

Key Words: *Machine Translation, Corpus, Translation, Purpose of Corpus.*

INTRODUCTION:

A corpus must be the true representative of a natural language. No one can deny that translation is one of the factors that have made the whole world a global village. As each speech community has a different language, there is a demand for translation. The advancement of technology has offered us the machine translation that is required nowadays. Corpus-based machine translation makes it easy for human in terms of time, money, and effort. The corpus-based machine translation may have problems in transferring the message accurately and faithfully from the source language into the target language. Translating cultural aspects and concepts is difficult for human translators, let alone the machine translators. As English and Arabic languages belong to different language families and have different cultures, we expect major linguistic and cultural problems that have to be tackled properly either by human translators or machine translators.

Types of Corpus:

There are different types of corpus and the followings are some of those types:

1. **Open vs. Closed:**
   An Open corpus refers to a corpus that receives more entries and undergoes constant expansion as done by the dictionary creators to add the newly coined expressions and to check the changes in the meanings that occur from time to time. The open corpus gets more entries from time to time, therefore, it gets larger and larger day in day out such as Longman Written American corpus and the bank of English.The closed corpus has a fixed size and never accepts new entries. It is also called static corpus by Baker as there is no change happens to it. There are many corpora that are closed and never get updated such as the manual dictionaries and some electronic dictionaries.
2. **Written Corpusvs. Speech Corpus**:
   Awritten corpus is a body of texts collected from different authentic sources such as printed books, published papers of different fields like science, arts, and etc. whereas,the speech corpus is a collection of speech audios including different varieties of conversations such as formal, informal, public, private, standard, dialectal and etc.
3. **Multimodal Corpus**:
   It is a collection of video data that is annotated along the timeline in order to code information in the video like gestural expressivity, emotions, or dialogue functions.
4. **Monolingual Vs. Multilingual Corpus**
   A monolingual corpus refers to a collection of texts of a single language that is used for the purpose of intralingual analysis. It can be used to analyze different genres of the same language, language teaching, and etc. A multilingual corpus refers to a collection of texts from two or more languages

and it is of two types as follows: comparable and parallel corpora. The comparable and parallel corpora are differentin terms of design to fulfill different purposes; typically the comparable corpus is used in contrastive studies, whilst the parallel corpus is used in translation researches (McEnery and Hardie, 2012: 20)

Benefits of Using a Corpus:

There are many benefits and functions of using corpora in the field of teaching, translation, or linguistics. Some of the benefits are mentioned below as follows:

1. **Correction:** The corpus can sometimes challenge our intuition, i.e. we, even as native speakers keep using something in a language, trusting our intuition that the words or grammar we use is common but the corpus analysis results that it is infrequent.
2. **Frequency:** Now, it is easy for all especially teachers and translators to check and find out which vocabulary or structure is frequently used. All this happens in seconds due to the virtue of corpus.
3. **Context:** Lexical items and grammatical structures are used in certain contexts and sometimes we cannot decide the exact context in which a certain word is used. Therefore, corpus can be approached for figuring out the matter.
4. **Collocation:** Words collocate with other words, and this is one of the challenges of that translators and teachers encounter. The corpus can save time and effort. The user of the corpus can just enter the parameter needed, and then the corpus software will show the co-occurrences.
5. **Language Teaching:** Teachers can greatly benefit from the use of corpus in their classrooms. Teachers need to teach vocabulary and their frequency of use, collocations, grammar,and etc. the use of corpus by teachers and students can make it as a shortcut to learn a language.
6. **Translation:** Translation is a challenging task for human beings as it takes too much time and effort to translation certain pages. The corpus software can be trained to translate from one language into another. There are a lot of electronic dictionaries and websites such as GOOGLE TRANSLATE that are used to perform translating. These machines are trained to mimic human translators and produce a translation with or without the assistance of human translators.

Machine Translation (MT)

1. Definition:

It refers to a situation in which softwareis used to translatea text or a speech from one language into another with or without the assistance of human. The machine translation is to assist human but not to replace them.

2. Importance of Machine Translation:

Computers and machines used for translation have made it much easy to translate from one language into another. In addition to that, they save time, money and effort.But, the questions that still roam in our minds are that: How accurate is MT? And Can MT replace the HT?

3. Machine Translation Process

We should do pre-editing for the input (the source text)and do post-editing for the output (the target language). The best machine software follows the following process: 1. Analyzing a source language and forming representation of the source language.2. Transforming representation of the source language into that of the target language form.3. Generating the target language representation. Machine translation can be bidirectional or unidirectional.

4. The Basic Approaches of Machine Translation

There are two main approaches of machine translation as follows:

1. **Rule-based approaches:** it is based on linguistic rules that are prepared and entered by linguists. It consists of three methods as follows:

a. Direct-based method: it is the oldest method and less popular. It can be used as unidirectional or bidirectional. It needs little syntactic and semantic analysis.

b. Transfer-based method: it is the best method so far as it follows three stages as follows: 1. The source text is converted into abstract source language-oriented representative. 2. The source language-oriented representative is converted into equivalent target language-oriented representative. And 3. The target language-oriented representative is generated.This method is the best as it consists of three dictionaries (source dictionary, target dictionary, and bilingual dictionary). It also consists of three grammars as follows: grammar for the source language, grammar for the target language, and grammar for transformation rules.

c. Interlingual-based method

Interlingual Machine translation: in this method, there is an intermediatory language, either natural or artificial, between the source and the target languages. This method is good in the case of multilingual

translation. If we need to translate from, for example, Arabic into Malayalam, we need to first translate from Arabic into an intermediatory language, be it English, and then from English into Malayalam.

2. Corpus-based approaches

Corpus-based Machine Translation: it refers to a huge collection of data and texts of both the source and the target languages. The parallel collected corpus is employed for the purpose of the translation. This approach has two further methods as follows:

a. **Statistical Machine Translation**: it depends on the statistical model. In this method, linguistic knowledge is not required for building the model. But it is difficult to create a massive parallel corpus. In this method, there are two models as follows: word-based and phrase-based. This model depends on the probability that is generated from the previously given parallel corpus.

b. **Example-based Machine Translation**: it is about the mapping between the SL and TL. This method is memory-based translation. It stores the input and then matches the sentences to be translated with what is already available in the system. If they match well, then the correct translation will be generated. This method depends on the concept of analogy.

Language and Culture

Any translator either a machine or a human one has to be bilingual and bicultural. Machine translators may be bilingual but not bicultural. Therefore, the translation produced by a machine or software is not accurate. There are two types of culture as follows: 1. Universal culture that is common to most or all languages in the world. 2. Specific culture that is related to a certain language but not to others. The machine may understand the universal culture but may fail to get the specific culture. Arabic and English have cultural differences as follows:

1. Environmental Differences

Arabic and English were born in different environments. Therefore, the translation of 'The news warms my heart' has a contradictory equivalent in Arabic as ' الخبر يثلج صدري'. The English 'warms' has been translated into Arabic as يثلج/ices which is the contradiction of 'warms'. This difference could be due to the fact that the English language was born in a cold environment where warm climate is the lovely weather for those countries where English was born. In contrary, Arabic language was born in a hot environment where the cold climate is the lovely weather for those countries where Arabic was born. The human translators can understand such a difference, whereas the machine translators may fail to capture such a difference.

2. Numbering Differences

The number 'nine' in the example 'The cat has nine souls' has to be translated into Arabic as 'seven' 'The cat has seven souls'. Another example is the number 'two' in 'A bird in the hand is worth two in the bush'. The number 'two' is translated into 'ten' in Arabic as عصفور باليد ولا عشرة على الشجرة/a bird in the hand is worth ten on the tree. To recognize such a difference is difficult for the human translators, let alone the machine translators.

3. Connotative Meaning

The connotative meaning depends on the culture of a certain speech community. The connotative meaning of the word 'moon' in Arabic is positive, whereas it is passive in English language. In Arabic, we describe the beauty of a woman by using the word 'moon' like 'she is like the moon'. Whereas in English, the word 'moon' has a negative connotation as seen in the following example 'she is like the moon', which means that she is temperamental or moody. Therefore, in English, we say that 'she is as changeable as the moon'. Another example is that the word 'owl' which has a positive connotation in English but a negative connotation in Arabic. In English, it is said that 'he is as wise as the owl'. In Arabic, describing a person with the word 'owl' means that the person is ill-omened. In Arabic, we say that 'as wise as Luqman' as'Luqman' is a symbol of wisdom but not the owl, in Arabic language. Such differences can be difficult for the machine translators to recognize and hence their translation may not be accurate. Human translators can manage understanding such differences.

4. Acronyms and Abbreviations

Translating acronyms and abbreviations is difficult for human translators, let alone the machine translators. Translating acronyms and abbreviations, especially the ones that are novel and have not been entered into the dictionaries, can create problematic issues. The machine and human translators need to be careful in translating those acronyms and abbreviations as accurately as possible. Human translators can search for the meaning of any acronyms and abbreviations in various sources. Whereas, machine translators fail to do so, therefore, their translation may not be accurate.

5. Novel Terms and Meanings

The machine translators need to be updated from time to time. Each language gets novel expressions and undergoes a semantic change from time to time. Therefore, it is important to translate those novel expressions as faithfully as possible. Those novel expressions and meanings should be added to the corpus of the machine translation so that the machine translators can recognize them and translate them accurately. If the corpus is not updated, the machine translation may not be faithful and accurate.The human translators can be updated on the daily basis as they interact with what is going on around them.

Conclusion

Machines and computers have been trained by the man to assist the man not to replace the man. There are many factors that play a crucial role in the quality and accuracy of translation.Some of these factors such as culture, pragmatics, physical context, and so on, cannot be obtained by the machine, at least so far. In the future, who knows, there may, again I say may, be a chance of improving the machine translators to overcome such problems but not to substitute the man.

References

- A Corpus-based Machine Translation Method of Term Extraction in LSP Texts, Theory and Practice in Language Studies, Vol. 4, No. 1, pp. 46-51, January 2014, © 2014 ACADEMY PUBLISHER Manufactured in Finland. doi:10.4304/tpls.4.1.46-51.
- Corpora in translation studies, by DOROTHY KENNY.
- Corpus-based Approaches to Translation Studies, APPROCHES FONDÉES SUR LES CORPUS DANS L'ÉTUDE DE TRADUCTION, SHENGuo.
- Corpus-based Machine Translation: Its Current Development and Perspectives, Zhou Dajun and Wang Yun, Department of Basic Courses, Naval Aeronautical and Astronautical University, Yantai, China.
- Corpus-based Translation Studies: A Quantitative or Qualitative Development?, Dorothy Kenny, Dublin City University, 2015.
- Goals and method of corpus-based translation, Cambridge University Press, 978-0-521-49957-6 - Corpus Linguistics: Investigating Language Structure and Use, Douglas Biber, Susan Conrad and Randi Reppen Excerpt.
- https://www.google.com/search?q=corpora+in+machine+translation&oq=corpora&aqs=chrome.3.69i57j69i60j69i59l2j69i61j69i59.5053j0j9&sourceid=chrome&ie=UTF-
- https://www.grin.com/document/385050
- https://www.sketchengine.eu/corpora-and-languages/corpus-types/
- Hunston S. Corpora in Applied Linguistics 2002
- Introducing corpora to translation studies, Maeve Olohan, Landon and New York.
- Machine translation and corpus-based methods, Cecilia Hemming, University of Gothenburg, 12 January 2015.
- Machine Translation in Europe, university of Geneva, 1227 Geneva Switzerland.
- rong.https://www.google.com/search?q=corpus+linguistics&oq=corpus+linguistics&aqs=chrome..69i57j0l5.15866j0j4&sourceid=chrome&ie=UTF-8.
- "Using corpora in Translation Studies: The state of the art", Ming Yue, Zhejiang University, 2019.
- What is a corpus and why are corpora important tools, Kristina Nilsson Björkenstam Computational Linguistics, Stockholm University.