TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 14368 - 14372

Machine Translation Models and its Challenges when Applied on Various Domain Corpus

K.Soumya[1], Dr.Vijay Kumar Garg [2]

[1]Research Scholar, Computer Science and Engineering, Department of Lovely Professional University, Punjab and Assistant Professor in VBIT, Hyderabad.

[2]Associate Professor, Department of CSE, Lovely Professional University, Punjab

Article Info
Volume 82
Page Number: 14368 – 14372
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 28 February 2020*

Abstract

Machine translation is an art of language engineering and is a sub branch of computational linguistics that carries out automated translation of human languages from one source language to another target language. There are many approaches to machine translation beginning from the basic existing statistical MT and rule based MT to the present neural machine translation approaches. Even though it is proved that SMT has noticeable less performance and accuracy than neural machine translation, it is also proved that SMT is still more suitable for different databases of text. So this paper would give a systematic review on various research done in machine translation before to neural machine translation.

Keywords:*Language engineering, Machine translation, computational linguistics, automated translation, statistical machine translation.*

I. INTRODUCTION

The only means by which human communicate with each other is language. and, all over the world there are above 4000 different languages used by various communities of people. And this shows the diverseness in languages all over the world. And it is difficult and impossible for every people to understand and communicate in all the languages. So here comes the importance of language translation between two different languages. Language translation with the help of a mediator or a third person who can understand both the languages is the traditional method applied before the invention of automated language translations done by machine.

So, Machine translation is a process of automated translation of languages from one to the other language with the help of a computer system. So, it is must that a complete automated system have a very sound knowledge of source and destination languages, in terms of subject matter and also about grammatical concepts.

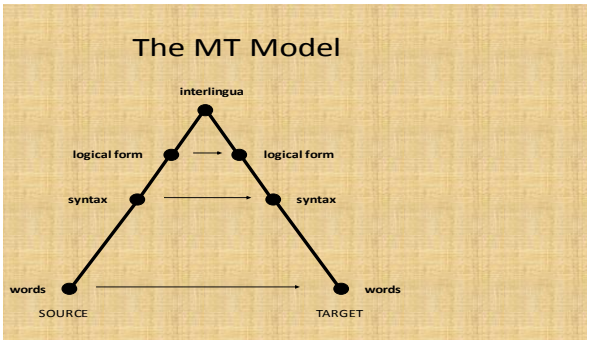

Fig 1: The MT model

II. HISTORY OF MT:

Language is both an art and science that converts one language to another language without losing any semantic meanings. MT systems are in existence since 1940s and has recently flourished due to the web's proliferation. MT was the first computer based NLP application and its history is very old. The field is said to have served as the computer science forcing function itself, due to the cold war in 1960s, the search for automatic means of translation between languages took on importance.

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193 - 4120 Page No. 14368 - 14372

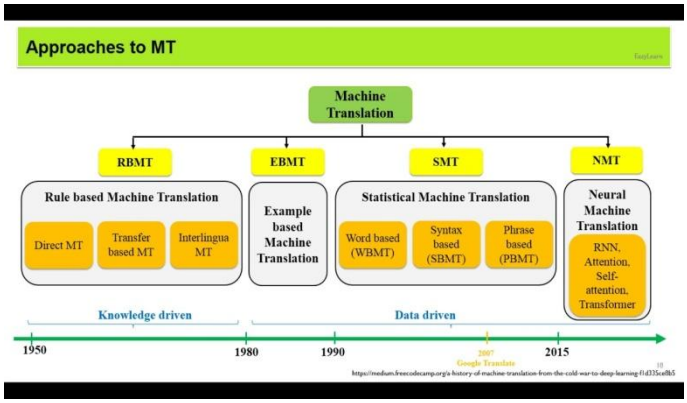

Fig 2: Approaches used for MT

III. MT APPROACHES:

A. Rule based approaches:

Different rule based approaches of MT are:

Direct MT:

This is kind of MT where if the input sentence have exactly the same match in the corpus or database of the stored sentences, then it is simply matched with the sentence and given as output.

Transfer Based MT:

In contrast to the direct machine translation, transfer based machine translation first does analysis on the source language text in order find out its grammatical structure, its phonological patterns and its semantics, and creates an intermediate form that can be used to transfer words of sentences from one language to the other.

Interlingua MT:

This is an alternative approach to direct and transfer based MT. The term interlingua indicates it's intermediate and abstract representation is neutral and has a capability of projecting source language and also representing a target language.

B. Example based MT:

It is basically a translation by correlation and relies completely on proportional analogy. In this, the MT system is given with source language database of text or corpus and also its target language corpus too. And correlation analysis is done to do translation between languages. i.e, it is assumes that if a previously translated sentence occurs again, then the same translation is likely to be happened. Which means that it uses trained translations and previous translations as examples for new set of sequences.

C. Statistical based machine translation:

Statistical machine translation:

This system does translation on text selected from bilingual parallel text databases or corpora. And its basic idea is to use a parallel corpus as a training set of translated text. Basically a simple SMT have a translation model and a language model. And these two models depend on calculating conditional probability of similarity of text or words in parallel corpora.

Word based machine translation:

In word based machine translation first the words of source language are mapped to the words of target language, and the permutations of words mapping are also calculated as there could be multiple similar words that may be mapped to words of target language, then the words are aligned so that it would give proper translation required.

Phrase based machine translation:

In contrast to the word based MT phrase based machine translation considers group of words or strings or phrase mapping between source and target languages.

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193 - 4120 Page No. 14368 - 14372

Table I: Pros and Cons of Statistical Machine Translation by considering challenges of SMT into consideration.

Statistical Machine Translation	
PROS	CONS
• It is easy to train and add new languages to statistical MT models when compared to other MT models. • SMT requires less virtual space than other models of MT. • It is easy to operate and train smaller systems with SMT. • If the model is well trained, a customized corpus can consistently translate the source content to a target content more accurate than NMT.	• It is difficult to analyze and translate a source content to a target content in SMT, if the training corpora doesn't include more feasible and similar content. • In this case, SMT may have a strong fall in accuracy. • SMT models depends on bilingual content. • SMT is expensive too, because it takes more time in preprocessing and corpus creation. • It is difficult to fix problems in model once it is implemented.

Syntax based machine translation:

This model of MT completely depends on grammatical and syntactical rules of both source and target natural languages. Where sentences are considered in the form of individual parts like verbs, prepositions, adverbs, conjunctions etc and reframed again and written as parts of sentences by following the grammatical rules of destination language.

Table II: Pros and Cons of Rule Based Machine Translation by considering challenges of RBMT into consideration.

Rule based machine translation	
PROS	CONS
• Based on morphological theories • Can be applied to languages with limited set of rules for forming language. • It is easy to verify errors with this method.	• Must have complete knowledge of linguistic rules of languages • Inconsistency problem occurs because there is less possibilities for translations. • It is expensive to maintain and improve the translations.

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193 - 4120 Page No. 14368 - 14372

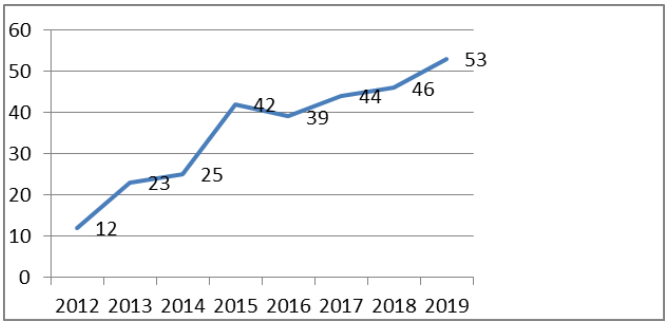

Fig 3: number of papers published and review (from 2012 to 2019).

Fig 3 illustrates number of journal articles in different duration of years in machine translation research area. Even through the initial search for articles resulted in a major number, there are more articles published from 2012 until 2019, but shown numbers are considered. Regardless, the rapid increase in the articles highlights the awareness and importance of this area among the academic community, practitioners, mainly in health care and even governments worldwide. Despite the variations and increase in the number of articles, this research domain is still emerging and also it appears that this research domain requires further in-depth conceptual as well as empirical and more analytical research studies.

Table III. Included databases

Database	Number of papers considered for final
PubMed & PubMed Central	52
CINAHL	11
ACM Digital Library	46
IEEE Xplore	51
MT Archive	65

Following fig 4 shows the categorization of reviewed papers into different publication type. As shown in the above figure it is clear that the majority of the publications are research papers, followed by literature review then general review and then based on view point. The more contribution of research papers in the review is clearly indicating the importance of machine translation area in different sectors (e.g. healthcare, government, and many more). And most of the authors concluded in their publications that, still there a need for improvement in the field of translation by applying it on various corpus.

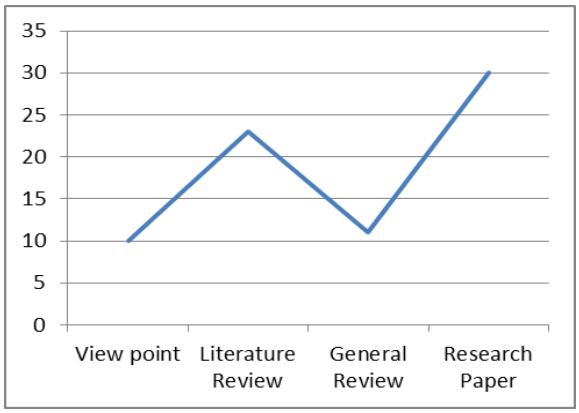

Fig 4: Classification of publication types

In the following Fig 5 shows statistics about review done on the MT algorithms applied on various domains. With the data presented in above figure it is clear that the majority of the publications are using statistical machine translation and neural machine translation techniques. And almost all the authors concluded in their publications that there is still a need for a lot of improvement in the field of translation in different domains.

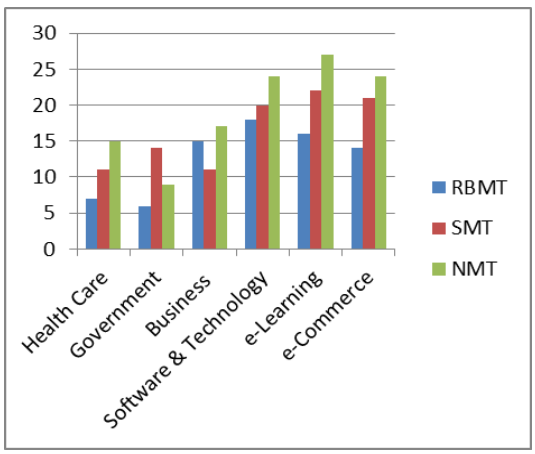

Fig 5: Review of MT algorithms when applied to various domains

Published by: The Mattingley Publishing Co., Inc.

January - February 2020
ISSN: 0193 - 4120 Page No. 14368 - 14372

IV: CONCLUSION

This paper have presented a holistic view about MT and their importance in different areas and application of different MT algorithms on different data bases of text or domain corpus. Based on the information discovered as the result of existing research studies, it would be easy for aspiring researchers to know about various advantages and disadvantages of applying different MT models on distinct kinds of databases of text or corpus of individual domains.

REFERENCES

[1] Turner AM, et al. (2014). A comparison of human and machine translation of health promotions materials for public health practice: time, costs and quality, Journal of Public Health Management and Practice, 20(5):523-9.

[2] Oladosu E, Esan A, Edayanju I, et al. (2016). Approaches to Machine Translation: A review.

[3] Journal of Engineering and Technology, 1(1):120-126.

[4] Weiss RJ, Chorowski J, Jaitly N, et al.(2017). Sequence-to-Sequence Models Can Directly Transcribe Foreign Speech. CoRR. abs/1703.08581.

[5] Turner AM, Desai L, Dew K, Martin N, Kirchoff K. (2015). Machine Assisted Translation of Health Materials to Chinese: An Initial Evaluation. MEDINFO, 979.

[6] Taylor RM, Crichton N, Moult B, et al. (2015). A prospective observational study of machine translation software to overcome the challenge of including ethnic diversity in healthcare research. Nursing Open,2(1):14-23.

[7] Liu W, Cai S, Ramesh BP, Chiriboga G, Knight K, Yu H. Translating Electronic Health Record Notes from English to Spanish: A Preliminary Study. ACL-IJCNLP 2015. 2015 Jul 30:134.

[8] Greg P. Finley, Erik Edwards, Amanda Robinson et al. (2018). An automated medical scribe for documenting clinical encounters, in Proceedings of NAACL-HLT, Association for Computational Linguistics, pages 11–15.

[9] DimitarShterionov, Riccardo Superbo, Pat Nagle et al. (Sep 2018). Human versus automatic quality evaluation of NMT and PBSMT.in Machine Translation , Volume32, Issue 3, pp 217–235.