A Survey on Machine Learning-based Translation Systems using Rule-based approach

Jasmine Kaur,
School of Computer Science and Engineering
Lovely Professional University.

Abstract: Machine translation is defined as the branch of artificial intelligence which covers the task of converting a source language to any other target language from the set of natural languages. The meaning of the text should be preserved and output should be fluent as well as correct. It is one of the applications of NLP (Natural Language Processing), which is the study of interaction between human languages and the computers. It is important for the computers to read, analyze, understand and derive meaningful information from the human natural language. There are various types of machine translation available. This paper discusses them in detail.

Keywords: NLP, Rule Based, Empirical, Machine Learning, SLR

Introduction

The benefits of using a machine translator instead of doing it manually are numerous. It can be time-saving and can optimize effort and money. It can also provide confidentiality and multi-lingual support which may not be possible in case of manual translation. However, problems may arise if context consideration is not taken into account. Similarly, languages may be ambiguous or have different structures. Also, machine translation is still not fully accurate. There are broadly two kinds of approaches that can be followed in order to design a machine translation system. They are as follows:

1. Language-based approach: Here, the language is chosen as a criterion for machine translation. The languages can be categorized as of Indian origin or non-Indian (foreign) based on their origin, history and current use.

2. Technique-based approach: Here, the technique employed for machine translation is opted as a criterion for categorization purposes. They are further classified as follows:

   (a) Rule-based approach: Here, we are concerned with the creation of rules which are derived using the syntactic, semantic and morphological analysis of the language. The human experts create rules in the form of corpus which are to be referred by the machine translator while processing any given input.

   (b) Empirical approach: This kind of approach makes use of source language information for carrying out the process of translation. This source information is available to the system in the form of training data which are nothing but the translation examples. This approach can be further classified as:

      i. Example-based approach: This approach takes into account the set of available examples for the translation process. These examples are saved in the form of bilingual databases where target language translation for the source language is presented. The translation system needs to map each and every input to the entries present in the database, which is a tree-like structure, and find the appropriate

      ii. Statistical-based approach: This kind of approach considers the probability distribution of sentences, words and phrases that are present in a certain language. This distribution is representing the probability of different ranges that a random variable can translate, achieve or a probability that a random variable
lies in a subset of that range.

LITERATURE REVIEW

A rule-based approach is used to achieve machine translation as described by Mishra et al. [1]. They have developed a prototype model for translation from English to Sanskrit. The proposed approach, made use of a set of transfer rules which were hand-written and converted the lexicons (could be paragraphs, sentences or phrases) from English language to Sanskrit language. Since Sanskrit is a morphologically rich language, instead of making most use of syntactical structures they used morphological markings to identify different parts of speech. The system works for different classes of input text and presents the results in GUI (graphical user interface) form. The performance evaluations for this system were compared with other machine translation methods such as bilingual evaluation understudy (BLEU), unigram precision (P), unigram recall (R), F-measure (F).

Bahadur et al. [2] designed a process engine named EtranS which accepted sentence in English and produced its equivalent sentence in Sanskrit after translation. The approach followed is rule-based and they discuss firstly some syntactical features of the Sanskrit language pursued by a comparison between Sanskrit grammar and Context Free Grammar. In addition, the grammatical contrasts between Sanskrit and English languages are discussed in brief. The database which was specifically created for the aforesaid translation process was created keeping in mind various properties of Sanskrit language. It had multiple tables viz. Vibhakti – roop, Phrase – break, Bi corpus, Dictionary, Vibhakti - kriya – number and Kriya - roop. The algorithm involved steps such as, breaking source sentence into tokens, gathering information related to semantics and syntax of the source sentence, look into the rule base for corresponding rules, for finding out the compatible words of the target language a mapping is done and finally the output is generated in the form of target sentence. The EtranS process engine comprises of 2 major elements. Saini et al. [3] have provided a review of various systems for Machine Translation (MTS) and conferred a consideration of the basic methodologies used by each one of them. Their survey mainly focuses on the current state of research in the machine translation field in India. They discussed different kind of approaches for machine translation along with the pros and cons of using each one of them.

Mishra et al. [4] provide a detailed view of example based machine translation system (EBMT). It is compared with RBMT and SMT systems. Various approaches of EBMT and related problems are discussed. The prominent characteristics of Sanskrit grammar and a comparison of English language and Sanskrit language is presented. The paper shows the divergence between English and Sanskrit language with the help of illustrative examples. It could be contemplated as representing the divergence between the order free classes of languages and the SVO (Subject-Verb-Object). The EBMT process begins by first matching the input to a set of examples and then extracting the corresponding translations. Then, the appropriate fragments are taken out from these mapped translations and at last these fragments are combined to generate the grammatically correct output in the form of target language. This procedure is carried out in the form of a two way process. Firstly, find out the fragments that correspond to the translation for the matched parts of the source text and secondly, perform their recombination accurately. The use of such systems generally occurs in the cases when the sources of information for a particular area of research are insufficient or unavailable due to some reasons.

Gupta et al. [5] use a rule-based approach to manifest knowledge representation of a machine translation procedure from Sanskrit language to English language. They generate Lexemes by making use of a parsing technique, which will further be used as inputs in the translation process. To produce the final output, a pattern based on dictionary and certain mapping rules have been generated. A Sanskrit sentence is presented to the first module as an input and its output is in the form of tokens i.e. input text is parsed and various root nouns and verbs are identified. This works as the input to the next (second) module which is the semantic mapper wherein the semantic word in Sanskrit is mapped to the corresponding semantic word in English. Since it is
easy to retrieve information through searching and analysis in a rule-based approach, using it is beneficial. On the contrary, as the procedure comprises of multiple phases the overall accuracy is dependent on the individual accuracies of every phase. A brief comparison of English and Sanskrit languages is carried out. Then, an algorithm for parsing the input sentence to generate Lexemes is presented. This is further used as input in the process of translation. Some rules for the mapping process and dictionary-based patterns are created to find out required output for generating final results.

Raulji et al. [6] discuss about various Machine translation systems which involve Sanskrit as either a source, target or a key support language. It also presents different techniques used by researchers for machine translation, such as, Corpus based, Rule based and Direct translation.

The principal objective of this paper is to find out the Sanskrit language suitability, morphology and apply most suitable MT techniques. They have discussed a system called ETSTS (English To Sanskrit Translator and Synthesizer) which is said to be based on either Rule-based or Example-based approach of MT. The system is able to further process and convert the translated target sentence into speech output form. The designed system is divided into multiple modules starting from input text, checking for grammatical and syntactical errors, tokenizer module, translation module, rule-based or example-based engine, bilingual dictionary, output text and at last, a waveform generation module.

Mishra et al. [7] combine an Artificial Neural Network (ANN) model with the conventional rule-based approach for machine translation system, which translates the input in the form of an English sentence (source language) into the output which is in the form of an equivalent Sanskrit sentence (target language). They have made use of a feed-forward Neural Network for choosing the Sanskrit words such as noun, adjective, verb, etc. from English language to a specific vector called Sanskrit User Data Vector (UDV). Since Sanskrit is a morphologically rich language, utilization of syntax is limited and sole morphological markings are used to discover different parts of speech, such as, Noun, Pronoun, Verb, Preposition, Adjective, etc. The results for translation are displayed in GUI format and more than one class of sentences are handled by the proposed system.

Shilon et al. [8] discuss a machine translation system which works with two languages, namely Hebrew and Arabic, in both directions. The outcomes of the aforementioned systems are also presented. Both Modern Hebrew and modern Standard Arabic language are actually Semitic languages which originated from the Middle-East. They share many similarities such as syntactic, semantic, lexical, orthographic, and morphological. However, they are not considered to be very close languages. Also, there is nothing like a parallel corpus for these two languages. Moreover, the linguistic properties of the two languages are discussed in brief. After that, the challenges that MT can face as a consequence of the dissimilarities between these two languages are presented. Some of these challenges are namely, lexical, morphological, syntactic challenges, etc. They have also shown how the approach to MT by using English as a pivot language is rejected. Instead they have chosen a new kind of approach which they call a linguistically-aware transfer approach. Finally, they have presented a set of simple examples and provided a tabular view of the evaluation scores plus the errors generated during MT process.

Shukla et al. [9] write about the translation divergence between the two languages of interest namely, Sanskrit and Hindi languages. Although, the two languages hold a lot of similarities including their origin and region of use, as both belong to the same Indo-Aryan family and Hindi has taken many of its features (lexicon and structure) from Sanskrit, both the languages have a great deal of divergence. The difference patterns are visible on the layer of some functional phrases and words such as Vibhaktis but are unobserved when it comes to using Dorr’s classification.

When we do not consider the cases where the Vibhakti is same in case of both the languages, and we only take up the condition where the cases are different for the two languages, the translation divergence can be divided into a total of 7 categories. It is also beneficial when we want to study vibhakti level differences between Sanskrit and other Indian regional languages.
Nguyen et al. [10] put forward an approach to design a system consisting of change rules for the task of Part-of-Speech (POS) tagging. The approach is based on a method of incremental information gathering and learning. Here, the transformation rules are stored in some predefined structure form and the addition of new principles takes place only when correction to the already existing rules is to be made. In addition, the interaction between the rules is allowed as although only in a controlled way, one rule can change the output of preceding rules. This approach is quicker when it comes to the training time and the tagging speed as shown by the experimental outcomes on nearly 13 languages.

Furthermore, it is compared with the other approaches. Firstly, the accuracy of TnT system and this approach using a lexicon-based initial tagger is compared. The accuracy gained by this approach reaches close to the Part of speech (POS) and morphological taggers which are best in their class. They further compared the accuracy of their model using TnT for the initial tagging (RDRPOSTagger+TnT) with the Mar-MoT model. It was found that TnT when compared to the earlier presented taggers, RDRPOSTagger and MarMoT, performs quickest when it comes to tasks of training and tagging. In addition, the RDRPOSTagger and MarMoT systems need almost identical time for training but when it comes to the speed of tagging the performance of RDRPOSTagger is far superior to MarMoT system.

Ebrahim et al. [11] presented statistical methods to improve the performance of English to Arabic language MT. They first discuss the MT problem with a brief history. After that, they present a brief description of SMT approach and give an overview of the SMT architecture. In addition, the problems that are faced because of the complexity of Arabic language are discussed. Various linguistic issues with Arabic language are studied and analyzed semantically, syntactically, orthographically and morphologically. To enhance the quality of Arabic to English language translation, certain changes are performed to the SMT in the name of preprocessing and post-processing. Preprocessing consists of modules such as, Morphological Tokenization, Syntactic Reordering, and Orthographic Normalization while post-processing involves modules such as, Morphological De-tokenization, Orthographic De-tokenization and Orthographic Enrichment.

Forcada et al. [12] have built up a free or open-source platform for rule-based machine translation which is named as Apertium. The platform is used to design a rule-based MT system with the help of a toolbox, an engine and the required data. The engine and tools for MT in Apertium are created by rewriting and extending two older MT systems, namely the SpanishCatalan MT system and the SpanishPortuguese MT system.

Both of these systems were developed by the Transducens group at Universitat dAlacant. An example for translation by Apertium is discussed followed by the limitations and work which can be done ahead. Linguistic data is present in the form of monolingual dictionaries. Compilers and different sorts of free or open-source devices in the Apertium stage are talked about. Finally, the interpretation results are accounted for some dialect sets created under this stage.

A system based on the usage of transfer-based approach that can translate a document from Hindi to English language is designed by Gehlot et al. [13]. The proposed system takes in text as input in the source language, checks its structure with the help of parsing and then the text is converted to target language using re-ordering rules. The transfer-based approach is found to be superior than Corpus-based MTS because Corpus-based MTS requires high number of word-aligned data for carrying out the translation which does not always exist for every language while the transfer-based MTS needs mere the knowledge about the features of both the languages of interest (that is, source language and target language) to make the transfer rules. It is claimed that completely accurate translations are generated for simple assertive sentences while for the complex and compound sentences, correctness reaches to certain accuracy only. At the end, some of the existing transfer-based MTS are compared. CYK (Cocke Younger Kasami) algorithm is used for the parsing purpose.

Hurskainen et al. [14] have designed an MTS for the task of English to Finnish translation, which uses rule-based approach. Generally, rule-based approaches make use of linguistic features such as lexicon and grammar of the involved languages. In the beginning, the source English language text is analyzed through
the process of parsing. A term for phrases is used; called Multi Word Expression (MWE). This MWE is kept separated before entering the lexical words for the target Finnish language. A major issue that is encountered is of Semantic disambiguation and focusing on how to deal with numbers in nouns is also crucial. There are two phases involved in the process of adding inflection tags. For the first phase, tagging of the major parts of the sentence is carried out, such as the verb, the noun consisting of subject, object, the indirect object, and different modifiers of the verb. In the second phase, tagging of the remaining parts of the sentence is performed, such as the adjective, pronoun, and number modifiers. The inflection tags given in the second phase depend on the tagging of the nouns performed in the first phase of the tagging process. After this, the stem boundary is marked. Then, conversion of inflection tags to surface forms takes place and at last, word ordering in target language is controlled with the help of POS tags.

CONCLUSION

The task of language translation holds great importance and performing it with an automated system is what makes it more efficient and time-saving. The field of natural language processing presents an interesting area of study as it deals with the way we humans communicate, in both written and spoken form. It tries to understand the interaction between the machines and the humans.

REFERENCES


