International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 06, 2020
ISSN: 1475-7192

MACHINE TRANSLATION: SANSKRIT TO ENGLISH

[1]NIKHIL RAMESH,[2]SHREYA KHANNA,[3]Ms. J BRISKILAL

ABSTRACT—*Translation is the key for communication among countries on a global scale. Machine translation is the process of translating the source language to the target language performed by a computer. Machine translation can be performed by applying a variety of techniques or approaches each of which has its own set of benefits and disadvantages.This paper proposes to perform machine translation from Sanskrit to English. The development of the machine translation system for Sanskrit, being an ancient language is a challenging task. Sanskrit is one of the oldest languages in the world and is now not widely in use, yet a large number of ancient texts are written in the language and hence a translation system will prove to be crucial in their translation and understanding. This paper proposes to incorporate the Paninian framework into a neural machine translation system*

keywords—*Machine translation, Paninan Framework, Neural machine translation*

I. INTRODUCTION

Machine Translation is the process of automated translation by a computer from the source natural language (such as Sanskrit) to the target natural language (such as English). Machine translation is an integral part of Natural Language Processing as well as in the Field of AI. There are in essence

3 types of Machine Translation systems Direct machine Translation, Rule based Machine translation and Corpus Based Machine translation.

The direct machine translation systems involve use of bilingual dictionaries for conversion from source to target language without any intermediate form. Rule based machine translation systems can be classified into 2 types transfer based approach and Interlingua based approach. Rule based systems involve the use of language rules for morphological, lexical analysis, semantic understanding and usually involves the use of some intermediate representations such as parse trees. The Interlingua based approach transforms the source language into an intermediary form and then this is converted into the target language, the advantage of this being that it is easier to extend to more than one language pair. The transfer based approach converts the source language into a source language specific intermediate form, then this is converted into a target language specific intermediate form and finally into the target language, this in unlike the Interlingua based approach where ideally the intermediate form in language neutral. The corpus based Machine translation systems learn from existing data how to translate between languages rather than using rules, it can be divided into 2 types statistical machine translation and example based machine translation.In statistical machine translation a statistical model is generated using existing data of translated text, and the model is then used to predict the most likely translation for a

[1] *Department of Computer Science, SRM institute of science and technology Chennai, India nikhilpbr97@gmail.com*

[2] *Department of Computer Science, SRM institute of science and technology Chennai, India shreyakhanna06@gmail.com*

[3] *Professor, Department of Computer Science, SRM institute of science and technology, Chennai, India, briskilal.j@ktr.srmuniv.ac.in*

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

given input sentence. In example based translation the system searches for sentences similar to input in the parallel corpora in real time and uses the similar sentences to perform translation.

Sanskrit is a language that has its roots in ancient India. It is one of the first documented languages in the world. India is a country with 22 official languages with Sanskrit being one among them. The state of Uttarakhand has declared Sanskrit as its second official language. A large number of ancient texts are in the Sanskrit language. Panini is a Sanskrit scholar who defined the rules that governed the Sanskrit language that was spoken at his time. The Paninian framework provided structure to the Sanskrit language and is still used in the translation of the Sanskrit language. The Paninian rules consists of four major components Astadhyayi (4000 grammatical rules), Sivasutras (Information with respect to phonological segments), Dhatupatha(A set of 2000 verbal roots), and Ganapatha(A list of 261 lists of lexical items)[1]. All this can be leveraged and applied to help create a machine translation system.

II. LITERATURE REVIEW

Subash C kak in [2] argues that the Paninian approach to defining a language could be critical in developing efficient computer based language understanding systems such as machine translation. The paper describes the Paninian grammar and draws similarities between the Paninian framework and machine translation systems. The Paninian rules such as the karaka describe the meaning of sentences by linking the actions of the verb to the agents and the situation, this is similar to the semantic analysis required as part of machine translation system. The paper summarizes that it should be possible to use Paninian style generative rules and Meta rules to describe most languages making machine translation easier.

Akshar Bharati et al in [3] describe how the karakas (semantico-syntactic relations) can be used to design a language accessor (anusaraka) which can act as a translation aid when dealing with any free word order language. The paper discuses how Paninian grammar uses vibhakti (inflectional) information for mapping sentences to semantic relations, and uses position information only secondarily. Thus with this it is possible to perform simple analysis of the source language (inflectional level) and generate sentences in the target language that may not be grammatically correct but is understandable to a person who knows the target language. The paper describes different parses that can apply this concept. The authors built and tested an anusaraka from Kannada to Hindi.

Vandan Mujadia et al in paper [4] describe a model to resolve entity pronoun references in Hindi dialogues based on the Paninian framework. The paper first discuses the various entity references or concrete references this, concrete references are noun phrases, quantifiers etc. Leveraging the semantico-syntactic related structures present in the Paninian framework to develop a rule based anaphora resolver. The resolver is found to give a 64% accuracy for dialogues between users and a 59% accuracy for a corpora that had play stories.

Akshar Bharati and Rajeev Sangal as part of their paper [5] focus on the use of the Paninian framework to develop parsers that apply to any free word order language. Since the meaning of sentences in free word order languages does not solely depend on position the karaka analysis(semantic- syntactic relations) part of the Paninian framework that describes the use of vibhakti to identify meaning and theta roles(the number and type of noun phrases required by a verb). The paper shows that a constraint based parser built on the Paninian Framework

reduces to a bipartite graph matching problem and provides a good parsing option for free word order languages and in comparison to the context free grammar based parser it does better in asymptotic time complexity.

In paper [6] authors Amita and Ajay Jangra discuss the application of Paninian Framework, with the focus on karaka analysis, to the English language. The paper discusses the karaka analysis and then speaks of the challenges involved in applying it to the English language. The main challenge faced is the structure of languages, since English is not free word order the structure of the sentence lends to the meaning, but the karaka analysis does not pay too much attention to the word. It concludes that the Paninian Framework can be used for the English language. Sai Kiran Gorthi et al in paper [7] discuss the use of karaka analysis in the development of a NLIDB (natural language interface to a database) system. They discuss a rule based approach as well as a statistical approach to develop the system. The general idea for both is to first obtain dependencies using the Stanford parser and then try to map these dependencies to six chosen karakas and finally use that to help develop the semantic meaning. The rule based approach had an accuracy rate of 52% and the statistical approach gave improved results in comparison(65-75 %). They were able to improve the overall performance of the NLIDB system

In paper [8] authors Namrata Tapaswi and Suresh Jain describe a rule based POS (Parts of Speech) tagger for the Sanskrit language. The algorithm used involves splitting the word into its root and suffixes and then applying a set of pre defined rules to find the POS for the particular combination of root word and suffixes. If the POS is not resolved in this step, it is left blank and in the next step context based rules are applied to obtain the POS. The algorithm was tested on 100 words of the language with 15 tags and returned 100%
accuracy.

In paper [9] and [10] Vaishali M Barkade and Prakash R Devale describe a rule based translation system from English to Sanskrit with 4 models lexical analyser, semantic mapping, translator and composition. Each model is defined using a set of rules that have been identified by studying the language of Sanskrit and English. The key parts of this paper are that the lexical Analyser uses a dependency grammar to identify relationship between the tokens and this info is used by the semantic mapper to match words from English to Sanskrit rather than being done word by word.

Sreedeepa H. S et al in paper [11] uses a rule based Interlingua approach, meaning the source language is tranformed into a intermediary form and then from this to the target language. The rules developed here were derived with the help of the Paninian Framework. Interlingua is represented using f-structure which gives the functional information about a sentence. The f-structure is then used to generate the English sentences. The system was tested for around 35 sentences and it gave accurate results for all the sentences.

Paper [12], written by Promila Bahadur et al, discuses the similarities between Sanskrit grammar and context free grammar. It describes a rule based approach to convert English sentences to Sanskrit. It uses Context free grammar for writing the production rules. The translation system consists of two components parsing and generator components. The parsing component generates tokens and identifies the grammatical information on the tokens. The unique part of the parsing component is the EtranS lexicon which consists of POS, unique id numbers that make it easier for mapping English to Sanskrit. The generator component is responsible for generating the output sentence in Sanskrit.

P Bahadur et al in paper [13] speak about a rule based machine translation system from Sanskrit to English and vice versa. They use a two way model to achieve translation in both directions. The key feature of this translation system is that during the syntax analysis it checks for grammatical correctness of sentences against its rules before proceeding to semantic analysis, if error is present an error message is returned.

Paper [14] by Vimal Mishra et al describes another rule based system for translation from English sentences to Sanskrit sentences. The rule based system used here however focuses on morphology to perform the POS tagging rather than using syntax rules since Sanskrit is a morphologically rich language. The evaluation of the system was done for randomly selected 20 English sentences using BLEU (BiLingual Evaluation Understudy), unigram Precision, unigram Recall, and F-measure results showed good levels of accuracy.

Paper [15], written by Sarita G. Rathod, describes a rule based and an example based machine translation system for conversion from English to Sanskrit. Both the systems are described in detail and implemented. The performance of both systems is compared using five different evaluation parameters precision, Recall, meteor, bleu and Fmeasure. It is found that the example based machine translation system has greater accuracy (10-12% is mentioned) than the rule based system for the English Sanskrit language pair

Sandeep R. Warhade et al in paper [16] describe a ubiquitous application that uses statistical machine translation system for translation of English to sanskrit It describes the use of a phrase based machine translation system containing 8 features. It consists of a language model, which estimates probabilities of a word string, thus helping in the POS tagging. A translation model is present which calculates the probability of all possible source and target sentence pairs thus choosing the target sentence with the best such probability.

Paper [17] by Vimal Mishra and R. B. Mishra discusses in detail, the example based machine translation system. This technique involves searching parallel corpora during real time. A sentence pair similar to the current input for translation is identified and this is used to learn how to perform translation for the current input. The sentence pairs in example base used in this paper each contain morphological info as well as root word correspondence between the Sanskrit and English sentences allowing greater accuracy.

Ganesh R. Pathak et al in paper [18] describe the transfer based machine translation system for English to Sanskrit. The transfer based approach involves having a separate intermediate form for the source language and one for the target language. The source text is tokenized, morphological analysis is performed, and the source sentence parse tree is created. Destination sentence parse tree is created, and the target sentence is generated.

Paper [19] by Vimal Mishra et al. describes a machine translation system to convert Sanskrit sentences to English sentences that combines the rule based approach with a neural network. A feed forward ANN is used in this instance and uses the morphology for POS due the morphological richness of Sanskrit language. The system is applicable to only 6 sentence types.

Paper [20] discuses the use of neural networks for creating a machine translation system. The idea involves training two models one using a recurrent neural network and the other using a LSTM (long short term memory network) model. Both models contain a decoder, which converts the source language into an intermediate form, and an encoder which converts the intermediate form into the target language.

Yonghui Wu et al discuss the Google neural machine translation system in paper [21]. This translation system uses Deep LSTM network models that consist of 8 encoder and 8 decoder levels. The system when tested

against evaluation parameters had a better result than current best levels for the language pairs tested thus showing the highest level of accuracy that had been achieved until that time.

Paper [22] discusses about Long Short Term Memory (LSTM) and the advantages of using it for language modelling. LSTMs offer larger context length when compared to feed forward networks and are easier to train than Recurrent Neural Networks; this makes them highly useful for Natural language processing applications.

Stephen Merity et al in paper [23] discuss the advantages of LSTMs as well the methods to optimise the LSTM for language models. The paper proposes the weight-dropped LSTM, a strategy that uses a DropConnect mask on the hidden-to-hidden weight matrices, as a means to prevent over fitting across the recurrent connections. The paper describes the optimisation techniques in using language modelling and believes it can be extend to other sequential problems as well

Paper [24], written by Ilya Sutskever et al., Discuses the use of LSTMs in dealing with sequential problems by testing it using a machine translation between franch and English since machine translation is a sequential problem. The system involved an LSTM to make an encoder and an LSTM to make a decoder. It was found to have a BLEU score of 34.8 on the WMT-14 dataset whereas the phrase based SMT system achieved a BLEU score of 33.3. This is attributed to the ability of LSTMs to deal with long term dependencies thereby making them highly suitable for sequential problems.

Kishore Papineni et al. in paper [25] describe an automated evaluation parameter known as BLEU for testing the accuracy of machine translation systems. This system in essence works by comparing the translations to reference translations and finding the number of matches, larger the matched greater the score. The BLEU evaluation of machine translation systems is found to be very close to the human accuracy in evaluation.

TABLE 1: SUMMARY OF THE PAPERS

TIT	AUTHO	MAIN IDEA	ADVANTAGE	DRAWBACKS
The Paninian approach to	Kak, Subhash	The paper suggests the use of rules similar to	N / A	N / A
Paninian framework and its application toAnusaraka.[3	Bharati, A., Chaitanya, V.,	Describes the use of karakas , a part of the Panini framework, to develop a language	It is less complex to develop than a complete machine translation system	A language accessor does not get the complete translation
Paninian grammar based	Mujadia, V.,	Describes a model to resolve entity	Since a rule based anaphora resolver	Has a accuracy of jus 64% for person
Parsing Free Word Order Languages in	Aks har Bhar	The use of the Paninian framework's karaka	N / A	N / A

An Annotation Scheme for English Language	] Amita and Ajay Jangra	Application of Paninian Framework, with the focus on karaka analysis, to the	It is shows the possibility of using karaka analysis to English making it	English has fixed word order and structure plays a main role in meaning while
Identification of karaka relations in an english sentence[7]	S. K. Gorthi, A. Palakurthi , R.	Describes the use of karaka analysis to improve the mapping of syntactic structures	The use of karaka analysis improves the overall accuracy	The rule based approach only has a 52% accuracy
Treebank based deep grammar acquisition and	Tapaswi, N., & Jain,	Describes a rule POS tagger for the Sanskrit language	Within the limited tags and dataset it	It has only been tested for limited words and tags and
English to Sanskrit Machine Translator: Lexical Parser[9]	P.R. Devale ,Ms.Vaishali M. Barkade	Describes a rule based translation system from English to sanskrit. The rules obtained by studying the two languages	It has a simple implementation and all the rules have been clearly stated	The accuracy though not specified is suggested to be average and not 100%
Interlingua based	.Sreedeepa	Describes a rule based	Showed a 100% accuracy for	Has only been tested for a

Translation[11]	Mary Idicula	the Paninian Framework	sentences	
Architecture of English to Sanskrit machine	Bahadur, Promila & Jain,	Rule based approach to convert English sentences to Sankrit. The rules are written	Context free grammar is easy to implement	Context free grammar does not work well with free word order
English to Sanskrit	Bahadur, P., Jain,	Discuses a two way rule based machine	N /	N /
English to Sanskrit machine translation	Mishra, V., & Mishra, R. B	Describes rule based system for translation from English sentences to Sanskrit	The BLEU scores for 20 random sentences ranges	The system was only tested against 20 sentences
English to Sanskrit Translator	Rathod S.G., Sondu	Describes a rule based and an example based	The example based system is found to have a	The rule based system has a accuracy around
English-to-Sanskrit Statistical Machine	Warhade, Sandeep R.,	Describe a ubiquitous application that uses statistical machine	Application was ubiquitous and provided good accuracy levels	The statistical models require large amount of computation to be
Study of Example based English to Sanskrit	Mishra, Vimal and Mishra, R.	Discusses in detail, the example based machine translation system. Which involves identifying a	Benefits being Computational cost and system building cost are low and The	One major drawback is the requirement for parallel corpora (hard for
English to Sanskrit Machine Translation	Ganesh R. Pathak Sachin P. Godse	Describes the transfer based machine translation system for English	Provides a greater accuracy then a	It is limited to a single language pair and difficult to extend to other
ANN and Rule Based Model for English to	Mishra, Vimal and Ravi Mishra	Describes a machine translation system to convert Sanskrit sentences to English	The system shows a good accuracy	It is limited to only 6 sentence types

On the Properties of Neural Machine	Cho, Kyunghyun, Bart van	Describes a machine translation system with an encoder and decoder built using	It was applied for an English – Sanskrit translation and	Both the RNNs and LSTMS struggle when the sentence length
	Bengio			
Google's Neural Machine Translation	Wu, Yonghui, Mike Schuster	Describes a translation system that uses Deep LSTM (long short	Acheived accuracy rates greater than the best accuracy	It has huge computational requirements to perform well
LSTM Neural Networ	Sundermeyer, Martin,	Discusses about Long Short Term Memory (LSTM) and the	It has larger context memory than	N / A
Regularizing and Optimizing LSTM Language	Merity, Stephen, Nitish Shirish	Discusses the advantages of LSTMs as well the methods to optimise	The model beat the existing custom built RNN cells by 1	The implementation of the LSTM can be
Sequence to sequence learning with neural networks[24]	Ilya Sutskever, Oriol Vinyals, and Quoc	Discuses the use of LSTMs in dealing with sequential problems by implementing and	It was found to have BLEU score of 34.8 , higher than the BLEU score of 33.3 of	It had little difficulty
BLEU: a method for automatic	K. Papineni	Discusses an automated technique for the	It saves a lot of time as compared to the	Despite its accuracy it can’t guarantee 100%

SUMMARY

From the literature Survey it can be inferred that the Paninian framework gives a complete description of the Sanskrit language and it can be used in the development of a machine translation system. A number of attempts to develop a machine translation system from Sanskrit to English have been made, with a large percentage of them being rule based approaches. From the approaches studied it can be understood that neural machine translation systems using LSTMs offer some of the highest accuracy rates for a machine translation system. The system proposed plans to use the Paninian framework to normalise the corpus and then train a LSTM based machine translation system

REFERENCES

1. Paul Kiparsky. 2009. on the architecture of Panini's grammar. In Gérard Huet, Amba Kulkarni, and Peter Scharf, editors, Sanskrit Computational Linguistics, Springer, Berlin, Heidelberg, pages 33–94
2. Kak, Subhash C.. "The Paninian approach to natural language processing." Int. J. Approx. Reasoning 1 (1987): 117-130.
3. Bharati, A., Chaitanya, V., & Sangal, R. (1994). Paninian framework and its application toAnusaraka. Sadhana, 19(1), 113–127.doi:10.1007/bf02760393
4. Mujadia, V., Agarwal, D., Mamidi, R., & Sharma, D. M. (2015). Paninian grammar based hindi dialogue anaphora resolution. 2015 International Conference on Asian Language Processing(IALP). doi:10.1109/ialp.2015.7451530
5. Akshar Bharati Rajeev Sangal "Parsing Free Word Order Languages in the Paninian Framework," ACL '93 Proceedings of the 31st annual meeting on Association for Computational Linguistics Pages 105-111, June 22 - 26, 1993
6. Amita and Ajay Jangra. "An Annotation Scheme for EnglishLanguage using Paninian Framework." (2015).
7. S. K. Gorthi, A. Palakurthi, R. Mamidi, and D. M. Sharma.Identification of karaka relations in an english sentence. Proceedings of ICON, 2014
8. Tapaswi, N., & Jain, S. (2012, September). Treebank based deep grammar acquisition and Part-Of-Speech Tagging for Sanskrit sentences. In Software Engineering (CONSEG), 2012 CSI Sixth nternational Conference on (pp. 1-4). IEEE
9. "English to Sanskrit Machine Translator: Lexical Parser " Prof P.R. Devale ,Ms.Vaishali M. Barkade, international journal on computer science and engineering(IJCSE) ISSN-0975-5462,Vol.02,NO 10
10. "English to Sanskrit Machine Translator Semantic Mapper " Prof P.R.Devale ,Ms.Vaishali M. Barkade. International Journal of EngineeringScience and Technology. 2.
11. Interlingua based Sanskrit-English Machine Translation "Ms.Sreedeepa ,Dr. Sumam Mary Idicula 2017 International Conference on circuits Power and Computing Technologies[ICCPCT]
12. Bahadur, Promila & Jain, Ajai & Singh Chauhan, Durg. (2015). Architecture of English to Sanskrit machine translation.

13. Bahadur, P., Jain, A., & Chauhan, D. S. (2011). English to Sanskrit machine translation. Proceedings of the International Conference & Workshop on Emerging Trends in Technology - ICWET'11.doi:10.1145/1980022.1980161
14. Mishra, V., & Mishra, R. B. (2012). English to Sanskrit machine translation system: a rule-based approach. International Journal of Advanced Intelligence Paradigms, 4(2),168.doi:10.1504/ijaip.2012.048144
15. Rathod S.G., Sondur S., "English to Sanskrit Translator and Synthesizer (ETSTS)", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 12, December 2012
16. Warhade, Sandeep R., Bharati Vidyapeeth and Prakash Devale. "English-to-Sanskrit Statistical Machine Translation with Ubiquitous Application." (2012).
17. Mishra, Vimal and Mishra, R. B., 'Study of Example based English to Sanskrit Machine Translation', Journal of Research and Development in Computer Science and Engineering, Polibits
18. Journal Article Ganesh R. Pathak Sachin P. Godse "English to Sanskrit Machine Translation Using Transfer Based approach" 2010 AIP Conference Proceedings
19. Mishra, Vimal and Ravi Mishra. "ANN and Rule Based Model for English to Sanskrit Machine Translation." (2010).
20. Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau and Yoshua Bengio. "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches." SSST@EMNLP (2014).
21. Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes and Jeffrey Dean. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." CoRR abs/1609.08144 (2016): n. pag.
22. Sundermeyer, Martin, Ralf Schlüter and Hermann Ney. "LSTMNeural Networks for Language Modeling." INTERSPEECH (2012).
23. Merity, Stephen, Nitish Shirish Keskar and Richard Socher. "Regularizing and Optimizing LSTM Language Models." CoRR abs/1708.02182 (2017): n. pag.
24. Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, pages 3104–3112, Cambridge, MA, USA, 2014.
25. K. Papineni et al. 2002. BLEU: a method for automatic evaluation of machine translation. In Proc. of the 40th ACL, pages 311–318.