iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45888

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Review on Machine Translation from English to Kannada

Pankaj Dwivedi[1], Shraddha C[2], Mahesh Mahaling Jottepagol[3], PrahladM Vijay[4], Rajashekaramurthy MC[5], Pruthvi J[6]
[1]*Educational Technology Unit Central Institute of Indian LanguagesMysore, India*
[2, 3, 4, 5, 6]*Department of Computer Science & Engineering Vidyavardhaka College of Engineering, Mysore, India*

Abstract: Interlingual is a machine translation tool that uses anartificial language to convey the meaning of reallanguages. The process of converting text from one language to another is known as machine translation.This study provides a better model of machine translation system for English-to-Kannada sentence translation that employs statistically based techniques. Here, we use Moses approach. Moses is a statistical machine translation system. systems are trained on huge amounts of parallel data as well as even bigger amounts of monolingual data in statistical machine translation. Parallel data is a set of sentences in two languages that are sentence- aligned, meaning that each sentence in one languageis matched with its translated counterpart in the other. Moses training technique takes the parallel data and infers translation correspondences betweenthe two languages of interest by looking for co- occurrences of words and segments. The two main components in Moses are the training pipeline and the decoder. The training pipeline consists of a set oftools that take raw data and convert it into a machinetranslation model. The Moses decoder determines thehighest scoring sentence in the target language that matches a given source sentence.
Keywords: SMT, GIZA++, Phrased basedtranslation, Preposition, BLEU

I. INTRODUCTION

Kannada is an official language spoken in Karnataka and spoken in other regions of India. Kannada is spoken and written by around 7 crore people in India. We found a huge body of Kannada literature stretchingback 4000 years The process of using a computer system to translate words from one language to anotheris known as machine translation. Statistical machine translation (SMT) is a machine translation tool that contrasts with rule-based techniques as well as example-based machine translation. The parameters of statistical models derived from the analysis of bilingual text corpora are used to perform translations. SMT systems aren't usually suited to a certain pair of languages. During the training phase, a statistical MT system generates mathematical models that are used to acquire linguistic knowledge. A parallel corpus of well-aligned parallel sentences is required to train an SMT system. With or without language inputs, a statistical MT system can be taught. Annotating a corpus is a difficult process for any MT system, especially for languages with limited resources like Kannada

II. LITERATURE REVIEW

'A Review on Machine Translation Systems in India'

Here they have conducted the survey on major machinetranslation developments in Indian Context such as Anglabharti, Anubharathi, Anusaaraka, AnglaHindi, MaTra and various projects here the used technologies are pattern governed methodology, 'Generalized Example Based (GEB) and Raw Example Based (REB)' to enhance the translation performance. In Anusaaraka ,the used technologies are language knowledge and domain specific modules[1].MaTra is a project funded by TDIL, a human-assisted translational system for English source languages to Indian target languages, based on a transfer technique and frame-like structures. An example-based English language to Hindi, Kannada, and Tamil languages as well as Kannada to Tamil translation system was developed under the guidance of Balajapally et al. (2006), where a set of bilingual dictionaries consisting of a sentence–word dictionaries as well as a phonetic dictionary that includes parallel corpora and its mapping is used for a corpus size of 50,000 words[18]. Vamshi Ambati and U Rohini presented a hybrid strategy for English to other Indian languages in 2007, combining EBMT and SMT approaches with little linguistic resources[19] . On the basis of the manual and a statistical dictionary developed using an example database consisting of source and target parallel sentences, as well as SMT tools, work is now being done to develop English-Hindi as well as other Indian language translation systems. Anuvadaksh is a project of the EILMT consortium, and it is based on a hybrid approach that allows translation of sentences from English to six other Indian languages[5]. It consists of a platform as well as technology-independent modules, and it helps the multilingual community, starting with domain-specific expressions in tourism and healthcare and gradually expanding into other domains. Tree- Adjoining-Grammar (TAG), analyse and Generate Rules (Anlagen), and Example-based MT are the technologies used in this system.

'A Comparative Study of English To Kannada' Baseline Machine Translation System With Generaland Bible Text Corpus[2]
The main approach used is 'Statistical machine translation'. SMT is a type of machine translation that employs machine learning techniques. The amount andscope of the data are critical factors in the accuracy of these systems. Data is continuously pre-processed in the SMT system. In SMT, a sentence can be translated from one language to another in a variety of ways, and this approach considers every sentence in the target language as a possible translation of the input sentence, but only sentences with a high probability areconsidered, which is done using the probability distribution function p(x/y), where y represents the source language sentence and x represents the target language sentence[20]. To systematically handle the challenge, language models (LM), translation models (TM), and decoders are used. Other than SMT, other approaches are i)use of morphological analysis and finding most occurred words in the language to determine the source language ii)rule based approach.
iii) EBMT Example Based machine translation is based on the idea of reusing the already translated examples. Example based translation involves three major steps -Example acquisition, Matching and Recombination[21]. iv) using comparable corpora and PBSMT (Phrase Based SMT) but it is shown that restricting phrases to linguistic phrases or statistically motivated phrases decreases the quality of translation. To improve efficiency, a large number of parallel corpora must be evaluated. v) Factored Machine Translation Systems is a technique in which a word has several representations in the target language and is merged with the target language using linguistic information, although pre-processing of the corpus is required. The experimental setup done in this survey is English is a Subject-Verb-Object (SVO) language, but Kannada is a highly verb endings and morphologicallyrich language. Kannada has a subject-object-verb structure, whereas English has a sub-verb-obj structure. Kannada sentences contain gender and case indicators between subject and verb to preserve coherence from the start to the final word, making it challenging to apply SMT to the English Kannada language pair. The main structural difference between English and Kannada is the significant distance between the subject and the verb. Compared to other Indian languages Kannada is morphologically richer with respect to inflection case and gender markers[2].

'Sense Disambiguation of Simple Prepositions inEnglish to Kannada Machine Translation'
In this research paper main focus is given to preposition. A preposition is a word that comes beforea "noun" to indicate how that noun is related to the other nouns and verbs in the phrase. Following a preposition, the noun (reference object) is in the accusative case and is regulated by the preposition. Words that begin prepositional phrases are also known as prepositions. The literature on preposition disambiguation is divided into three categories. i) The preposition's semantics lexical. ii) The verb and the prepositional phrase (PP) that the verb uses as an argument. iii) The PP's head noun. Rather than evaluating each of them separately, combining them yields a decent outcome. The sense of the preposition has been determined by combining head (modified) and complement (modifier) information [23 ,24].The modified (head) is the phrase's head to which the PP affixes itself. The head noun of the PP is the modifier (complement). Based on [25] work, this study provides a strategy for disambiguating English common simple prepositions such as at, for, in, on, to, and with in the context of English to Kannada MT. The procedures fordisambiguating basic prepositions in English are as follows. i) Obtaining the proper parse. ii) Extraction ofcontext and semantics. iii) Selection of senses. Obtaining the Appropriate Parse, The prepositionalphrase (PP) data can be utilised to determine a preposition's meaning. The right parsing of the PP aidsin the selection of the correct sense. A phrasal verb is made up of a verb plus a particle (adverb or preposition) that has a distinct meaning than the component verb. It should not be translated just on thebasis of its component verb[16]. Because the constituents (verb and particle) work together to provide a specialised context-specific meaning that cannot be inferred simply from the original meaning ofthe constituents (verb and particle). The task of preposition disambiguation will need the identification of phrasal verbs. In the next step, The current study considers context as a collection of features that mightbe syntactic, lexical, or both. Syntactic context can be a changed relation, and lexical characteristics can include morphological information like a verb's TAM (Tense, Aspect, and Modality), class, category, and, inrare situations, the lexical item itself. WordNet and dictionaries are used to capture the semantics of modifier and modified. Hypernyms of a word may be found in WordNet[10] . We may quickly acquire the larger, more comprehensive class/concept for a modifier/modified by utilising this attribute. The WordNet's noise can sometimes lead to surprise and unexpected sense selection. We can get around this obstacle by looking it up in a dictionary. In Kannada, prepositions are translated as suffixes to the PP's head noun or as post-positions. A set of rules that are appliedin a linear fashion. These rules were created by hand and are kept in a rule file. The rule file now contains rules for the six typical basic prepositions mentioned before. On the basis of rules provided in a rule file, many Kannada meanings for a particular English preposition are chosen. The algorithm proposed is consisting of five steps,

1) Reading the input sentence.
2) Sentence pre-processing step.

3) Constructing I- tuple.
4) Disambiguation step.
5) Outputting the R- tuple's field6[4].

The system accepts a preposition- containing input sentence S and sends it to the sentence pre-processor, which performs pre-processing tasks such PP attachment extraction and phrasal verb identification before generating a new sentence S1. The I-Tuple function Object() { [native code] } creates an I-Tuple record from the pre-processed phrase S1. The Disambiguator looks for precise matches between I- Tuple records and each and every record in the rule file(R-FILE). If it discovers a match, the right Kannada equivalent sense for an ambiguous English prepositionis stored in the sixth field of the matched record. As a result, it provides the corresponding meaning of the sixth field of a matched record. During disambiguation, the module consults the Dictionary and WordNet databases.

'Kernel Method for English to KannadaTransliteration'

Machine transliteration is the process of converting a character or word from one alphabetical system toanother. It's a phonetic and orthographic conversion technique. It's useful in natural language applications like information retrieval and machine translation, especially when dealing with proper nouns and technical phrases, as well as cross-language applications, data mining, and information retrieval systems. The Support Vector Machine (SVM) is a binary classification machine learning technique. It's been used to solve a variety of real-world challenges, including Natural Language Processing. With maximum margin, SVM learns a linear hyperplane thatdivides the set of positive instances from the set of negative examples (the margin is defined as the distance of the hyperplane to the nearest of the positiveand negative examples). In terms of generalisationbounds for the induced classifiers, this learning bias hasbeen shown to have favourable qualities. The whole model has three important phases i) Pre-processing phase ii) Training phase using SVM iii) Transliteration phase which generates Kannada transliterations for a given English name. In pre-processing stage, SVMrequires that the training file be translated into a certainformat. The source language names are segmented and aligned with the corresponding segmented target language names during the pre-processing step. We constructed a parallel corpus of 40,000 Indian place names in order to train the SVM. The following is a listof the steps involved in pre-processing[11] In training stage, The corpus is converted into SVM input file format during the preparation step, with aligned sourceand target language names supplied as input and label sequences for training. The transliteration model istrained using this file. Finally in transliteration stage, Alist of English terms to be transliterated is compiled. These words are transformed to SVM test file format and then transliterated using the trained model, which returns the top N Kannada terms. The drawback of thismodel is, We used the SVM kernel to solve the problem of transliterating English into Kannada. The sequence labelling approach is used to mimic the transliteration system. When compared to other state-of-the-art machine learning algorithms, the framework based on data driven technique and one to one mapping approach simplifies the transliteration system development procedure and permits better gains intransliteration accuracy. The model has been trained on40000 words that include the names of Indian locales. The best 5 transliterations are used to assess the model. According to the results of the experiment, using the top five transliteration results greatly improves total transliteration accuracy. We anticipate that this will beextremely valuable in natural language applications, such as bilingual machine translation, and in a variety of other fields.

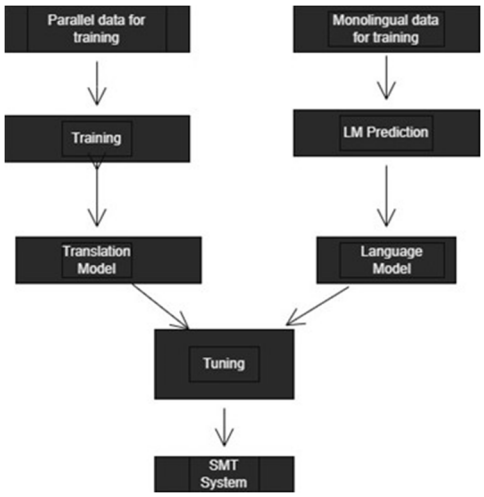
'COMPARATIVE STUDY OF FACTORED SMTWITH BASELINE SMT FOR ENGLISH TO KANNADA'

In this paper, The two approaches to develop Natural language processing are Rule based and Statistical based methods. A Rule based system requires modelling the language rules into a computer understandable format with mathematical models and algorithms. Process of annotating a corpus is a challenging task for any machine translation system particularly the resource poor language like Kannada. Factored machine translation captures the linguistic information that is the surface order of parts of speech for a given sentence and gives the output in the given pattern structure. Factored machine translation ensuressyntactic and semantic arrangements of the words in machine translation sentences. In the Kannada language corpus file, which may be found on the IIT Bombay website, using Indic tokenize. The Moses toolset will not normalize Kannada tokens since they are very agglutinative. To create a parts of speech tagged corpus for English and Kannada sentences we followed some steps. i) Creating parts of speech tagged corpus ii) Create Language model for Kannada iii) Training Factored Machine Translation iv) Decoding v)Testing[11]. We utilized the Stanford parser to analyse the Kannada Factored Corpus file in order to retrieve parts of speech tag sets. We constructed the corpus in lemma, derivatives, and parts of speech terms, which is the same format as the English factored corpus. We needed to build language sentences in order to establish a statistical machine translation system fordelivering language. In this study, we built a Kannada language model. For a given corpus, the languagemodel calculates the number of unigrams, bigrams, andtrigrams, as well

as their probabilities. We simplyrequire a surface language model for baseline machinetranslation, and we need a parts of speech language model to extract the arrival pattern of parts of speech tags in a given language phrase for factor machine translation. The annotated sentences in the factored corpus contain details such as lemma words, derivativewords, and parts of speech tags for both source and target language sentences. The mapping between lemma to lemma and their associated parts of speech isentered as 0-0, 2 in the training script parameter. For decoding a specific input sentence or file, the decoder uses the result of the training phase, such as a configuration file called moses.ini. To test a system in factored machine translation, we employed a test corpus as well as random sentences. The input should be delivered in factored sentences, and the outcome will be a regular surface sentence. We're training and testing the algorithm with a parallel corpus of 2500 sentences. The Kannada language has a total of 8706 words and 2466 unique terms, giving this English to Kannada translation a higher BLEU score. The total number of words in the English corpus is 12069, including 1609 unique terms. We utilize a parts of speech tagger and a shallow parser provided online from the IIIT Hyderabad LTRC website to annotate Kannada texts. We need to annotate the corpus with root words, derivative words, and parts of speech data separated by a delimiter such as the | sign to train the factoring system. The baseline static machine translation (SMT) results were taught using the Mosestool kit, and the baseline system was assessed using theBLEU score on sentences selected at random from thetraining set, rest set, and test set. In conclusion, We experimented with two approaches of statistical machine translation for English to Kannada languages as part of a project to develop language processing tools for Kannada. The trials on English Kannada corpora with and without annotated corpora reveal thatfactored machine translation improves the BLEU scorewhen compared to baseline machine translation for thesame corpus. With a small dataset of 3000 phrases, wewere able to achieve a 25% improvement in machine translation performance. To increase the performance of factored SMT, we can expand our trials with additional phrases in a factored corpus

III. METHODOLOGY

The machine translation using Moses can be doneusing two ways: i) Using parallel data. ii) Using Monolingual data.



Flow diagram of SMT using Moses

The steps involved in machine translation using mosesapproach is:

1) Preparing the corpus: We require parallel data (text translated into two distinct languages) that is aligned atthe sentence level to train a translation system.

The following procedures must be completed in order to prepare the data for training the translation system:

a) *Tokenisation:* This entails inserting spaces between (for example) words and punctuation.

b) *Truecasing:* Each sentence's first words are translatedto their most likely casing. This aids in the reduction ofdata sparsity.

c) *Cleaning:* Long and empty sentences, as well asplainly misaligned sentences, are deleted since they might cause difficulties with the training pipeline.

2) Language Model Training: The language model (LM) is created with the target language in mind to ensure fluent output. IRSTLM also offers a binary format, which Moses understands. For further details, see IRSTLM's documentation. The command line choices are fully explained in the KenLM manual,however the manual will help to design an acceptable 3-gram language model.

3) Training the Translation System: The most importantly, the translation model must be trained. To do this, we use a single command to perform word alignment (using GIZA++), phrase extraction and scoring, lexicalised reordering tables, and Moses configuration file creation. Using two cores on a powerful laptop, the translation took roughly 1.5 hours(Intel i7-2640M, 8GB RAM, SSD). This ini file specifies a model to decode (i.e. translate), however there are a few issues with it. The first is that it takes along time to load; however, this may be remedied by binarising the phrase table and reordering table, i.e. compiling them into a structure that loads rapidly. Thesecond issue is that Moses' weights for comparingmultiple models aren't optimised; if you check at the moses.ini file, you'll notice that they're set to default values like 0.2, 0.3, and so on.

4) Tuning: This is the most time-consuming portion ofthe procedure because it might want to preparesomething to read while it's going on. Tuning need a tiny bit of parallel data that is distinct from the trainingdata, therefore we'll once again get some data from WMT. It will run a lot faster to run in multi-threaded. The end result of tuning in an .ini file with training weights

5) Testing: It takes at least a few minutes for thedecoder to come up to speed. We can binarise the phrase-table and lexicalised reordering models in orderto get things started quickly. It can dine by creating suitable directory and binarzing the model. The translation is poor but comprehensible; keep in mind that this is a small data set for broad domain translation. Also, because the tuning process is non-deterministic, your results may vary significantly. To do so, we employ a different parallel data set from the ones we'veused previously. The trained model may then be filtered for this test set, retaining only the elements required to translate the test set. This will significantlyspeed up the translation process.

IV. CONCLUSION

We attempted to quickly discuss the various existing approaches to developing MT systems. Because Kannada language is morphologically rich in features and agglutinative in nature, most existing Indian language MT projects are based on a statistical and hybrid approach, according to the survey. This has encouraged researchers to choose these approaches to dealing with creating MT frameworks for Indian languages. SMT is employed in real time in (Google and Bing) free online translator tools for word-to-wordtranslation, and it continues to expand language options, but its efficiency is not up to par when it comesto sentence-to-sentence translation. Literature shows that the rule based machine translation process is extremely time consuming, difficult and failed to analyse accurately a large corpus of unrestricted text. The motivation for using SMT is to take advantage of the SMT system's robustness and linguistic knowledge of morphological analysis, and to systematically address the problem with the use of large volumes of bilingual corpora using a system combination approachand usage of language model, translation model, and decoder will increase the efficiency of sentence tosentence translation compared to other approaches. English corpus we see more no. of repeated words in each class , thus the no. of tokens are less and because of high frequency of words there perplexity factorwhile assigning the probability . Kannada has more number of unique words and thus the number of tokensare more and frequency is less compared to English language, therefore perplexity factor will also be less. Moses can be easily adoptive to any pair of language because of its phrased based and factored models for training the data. So, It’s best to use Moses for machine translation from the English language to Kannada language.

REFERENCES

[1] Maheshwari, Shikha & Saxena, Prashant & Rathore, Vijay Singh. (2019). A Review on Machine Translation Systems in India: Proceedings of ICETEAS 2018. 10.1007/978-981-13-2285-3_3.
[2] S. Saini and V. Sahula, "A Survey of Machine Translation Techniques and Systems for Indian Languages," 2015 IEEE International Conference on Computational Intelligence & CommunicationTechnology, 2015, pp. 676-681, doi: 10.1109/CICT.2015.123.
[3]] Km, Shivakumar & Namitha, B.N. & Nithya, R.. (2015). A comparative study of english to kannada baseline machine translation system withgeneral and bible text corpus. 10. 30195-30202
[4] . Parameswarappa and V. N. Narayana, "Sense disambiguation of simple prepositions in English toKannada Machine Translation," 2012 InternationalConference on Data Science & Engineering (ICDSE), 2012, pp. 203-208, doi: 10.1109/ICDSE.2012.6282320.
[5] Bandyopadhyay S (2004) ANUBAAD—Thetranslator from English to Indian languages. In: Proceedings of the VIIth state science and technology congress. Calcutta, India, pp 43–51
[6] M. M. Kodabagi and S. A. Angadi, "A methodology for machine translation of simple sentences from Kannada to English language," 2016 2nd International Conference onContemporary Computing and Informatics (IC3I), 2016, pp. 237-241, doi:10.1109/IC3I.2016.7917967.
[7] Antony, P. & Ajith, V. & Kp, Soman. (2010).Statistical Method for English to Kannada Transliteration. 70. 356-362. 10.1007/978-3-642- 12214-9_57.
[8] Godase, Amruta & Govilkar, Sharvari. (2015). Machine Translation Development for IndianLanguages and its Approaches. International Journal on Natural Language Computing. 4. 55-74.10.5121/ijnlc.2015.4205.
[9] Ganesh, Surya & Yella, Sree & Pingali, Prasad & Varma, Vasudeva. (2008). Statistical Transliteration for Cross Langauge Information Retrieval using HMM alignment and CRF.
[10] K. Sahoo and V. E. Vidyasagar, "Kannada WordNet - a lexical database," TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, 2003, pp. 1352-1356 Vol.4, doi: 10.1109/TENCON.2003.1273137.
[11] K. M. Shivakumar, N. Shivaraju, V. Sreekanta and

D. Gupta, "Comparative study of factored SMT with baseline SMT for English to Kannada," 2016 International Conference on Inventive Computation Technologies (ICICT), 2016, pp. 1-6,doi: 10.1109/INVENTIVE.2016.7823217.

[12] A. P.J., A. V.P. and S. K.P., "Kernel Method for English to Kannada Transliteration," 2010 International Conference on Recent Trends in Information, Telecommunication and Computing, 2010, pp. 336-338, doi: 10.1109/ITC.2010.85.

[13] Reddy, Mallamma. (2022). ENGLISH TO KANNADA/TELUGU NAMETRANSLITERATION IN CLIR: A STATISTICAL APPROACH.

[14] S. A. Angadi and M. M. Kodabagi, "A Robust Segmentation Technique for Line, Word and Character Extraction from Kannada Text in Low Resolution Display Board Images," 2014 Fifth International Conference on Signal and Image Processing, 2014, pp. 42-49, doi:10.1109/ICSIP.2014.11.

[15] S.N. Sridhar, Modern Kannada Grammar,
Manohar Publications & Distributors, 2007

[16] Yukiko Sasaki Alam, "Decision Trees for Sense Disambiguation of Prepositions: Case of Over," InHLT/NAACL-04, 2004.

[17] [Koehn, P.: MOSES a Beam-Search Decoder for Factored Phrase-Based Statistical MachineTranslation Models User Manual and Code Guide. University of Edinburg, UK (2009)

[18] Naskar S, Bandyopadhyay S (2005) Use ofmachine translation in India: current status. AAMTJ 25–31

[19] Sinhal RA, Gupta KO (2014) A pure EBMT approach for English to Hindi sentence translation system. I J Modern Educ Comput Sci 7, 1–8.
Published Online July 2014 in MECS

[20] Unnikrishnan P, Antony P J, Dr. Soman K P "A Novel Approach for English to South Dravidian Language Statistical Machine Translation System"

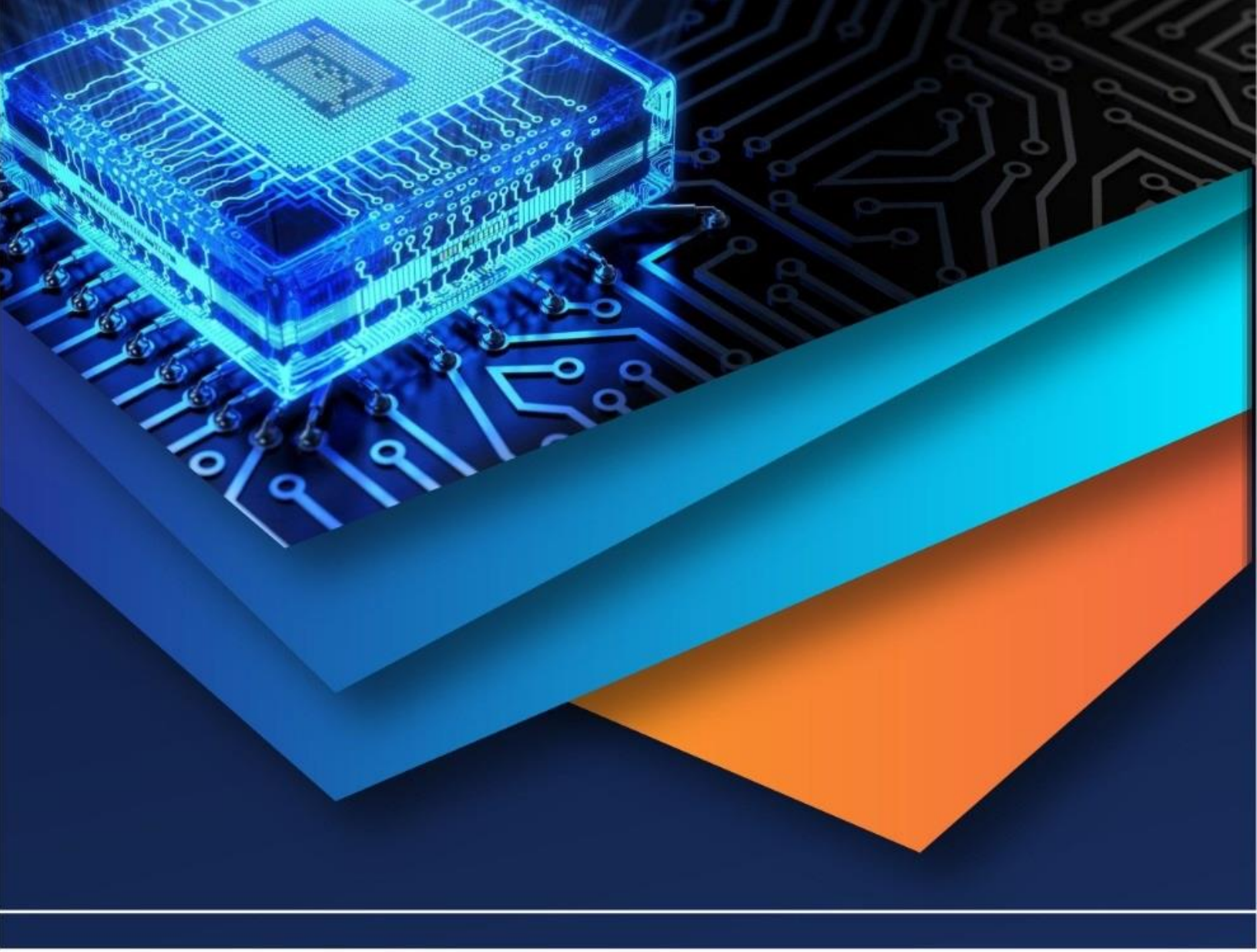
[21] Mallamma V Reddy, Dr. M. Hanumanthappa "Natural Language Identification and Translation Tool for Natural Language Processing" Department of Computer Science and Applications, Bangalore University, Bangalore, INDIA.

[22] David Chiang. 2007. Factored based translation.
Comput. Linguist., 33(2):201–228, June

[23] Patrick Saint-Dizier and Gloria Vazquez, "A compositional framework for prepositions," in IWCS4, Tilburg, Springer, lecture notes, 2001, pp.165-179.

[24] Yukiko Sasaki Alam, "Decision Trees for Sense Disambiguation of Prepositions: Case of Over," InHLT/NAACL-04, 2004.

[25] amar Husain, Dipti Misra Sharma and Manohar Reddy, "Simple Preposition Correspondence: Aproblem in English to Indian language Machine Translation," Proceedings of the 4th ACL- SIGSEM Workshop on Prepositions, ACL. Prague,Czech Republic. 2007.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY