IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-12, May 2019 E-ISSN: 2347-2693

Language Translator using Dictionary and APIs from English to Sindhi

Pinky Gangwani[1*], Samir Ajani[2]

[1,2]Department of Computer Science& Engineering, Jhulelal Institute of Technology, Nagpur, India

[*]*Corresponding Author: pinkygangwani08@gmail.com, Tel.: +91-9370219797*

Available online at: www.ijcseonline.org

***Abstract*—**In India Language Translation systems have been developed for translation from English to Indian Languages and from regional languages to regional languages. These systems are also used for teaching machine translation to the students and researchers. Most of these systems are in the English to Hindi domain with exceptions of a Hindi to English and English to Kannada machine translation system. English is a SVO (subject-verb-object) language while Indian regional languages are SOV (subject-object-verb) and are relatively of free word-order. A survey of the machine translation systems that have been developed in media for translation from English to Indian languages and among Indian languages reveals that the machine translation software is used in field testing or is available as web translation service. We present an analysis regarding the performance of the state-of-art Phrase-based Language Translation (LT) on Indian Sindhi languages. We report baseline systems on Sindhi language pairs. The motivation of this study is to promote the development of SLT and linguistic resources for these language pairs, as the current state-of-the-art is quite bleak due to sparse data resources. The success of an SLT system is contingent on the availability of a large parallel corpus i.e. Dictionary. Such data is necessary to reliably estimate translation probabilities.

***Keywords*—**Statistical Language Translation (SLT), Phrase-based Translation, Parallel Corpus, Natural Language Processing (NLP), Sindhi Phrase Word By Word.

I. INTRODUCTION

Language Translation (LT) can be defined as an automated system that analyses text from a Source Language (SL), applies some computation on that input and produces equivalent text in a required target language (TL) ideally without any kind of human intervention. It is one of the most interesting and the hardest problem in the field of NLP.

We will develop a dictionary of vocabulary and grammar rules from English to Sindhi in MS.SQL Server. We will develop an app that will translate inputted English phrases into Sindhi by scanning through dictionary and programming language API's. As much as comprehensive is dictionary the most accuracy will be shown. In short, Accuracy isbased on training dictionary. Higher Training higher accuracy.

The common approaches to machine translation are the rule-based approach and corpus-based approach.

- In the rule-based approach, the text in the source language is analyzed using various tools such as a morphological parser and analyzer and then transformed into an intermediate representation. A set of rules are used to generate the text in target language of this intermediate representation. A large number of rules are necessary to capture the phenomena of natural language. These rules transfer the grammatical structure of the source language into target language. As the number of rules increases, the system become more complicated and slower to translate. Formulation of a large number of rules is a tedious process and requires years of effort and linguistic analysis.
- In the corpus-based approach, large parallel and monolingual corpora are used as source of knowledge. This approach can be further divided into statistical approach and example-based approach. In the statistical approaches, target text is generated and scored through a statistical model, the parameters of which are learned from parallel corpus.
- The example-based translation approach is based on analogical reasoning between two translation examples. Example-based translation is essentially translation by analogy. An EBMT system is given a set of sentences in the SL (from which one is translating) and their corresponding translations in the TL, and uses those examples to translate other, similar source-language sentences into the TL. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again. EBMT systems are attractive in that they require a minimum of prior knowledge; therefore, they are quickly adaptable to many language pairs.

II. LITERATURE SURVEY

To have through study over research we have studied some approaches, methods and some machine translation basics.

We researched that, in a large multilingual society like India, there is a great demand for translation of documents from one language to another language. Most of the state government works in there provincial languages, whereas the central government's official documents and reports are in English and Hindi. In order to have an appropriate communication there is a need to translate these documents and reports in the respective provincial languages. Natural Language Processing (NLP) and Machine Translation (MT) tools are upcoming areas of study the field of computational linguistics. Machine translation is the application of computersto the translation of texts from one natural language into another natural language. It is an important sub-discipline of the wider field of artificial intelligence. There are certain machine translation systems that have been developed in India for translation from English to Indian languages by using different approaches. It is this perspective with which we shall broach this study, launching our theme with a brief on the machine translation systems scenario in India through data and previous research on machine translation.

We studied that An EBMT (Example Based Machine Translation) system requires a set of sentences in the source language and their corresponding translation in the target language. A bilingual dictionary comprising of sentence-dictionary, phrases-dictionary, words-dictionary and phonetic-dictionary is used for the machine translation. Each of the above dictionaries contains parallel corpora of sentence, phrases and words, and phonetic mappings of words in their respective files. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to occur again. A sentence may be seen as a combination of phrases. To translate, each sentence is divided into its constituent phrases and words, and these smaller units are translated by looking up in the sentence, phrase and word dictionaries. For words whose translation is not found, at least their phonetic translation (transliteration) is shown in the target language.

Rules of translation have been created, that allow substitution of a given noun with another noun, a verb with another verb and so on, without the need to enter every combination separately in the phrase database. This has been observed to improve the results dramatically. The phrase translations and phrasal rules play a significant role in this translation system. The advantage of this simple "good-enough translation" system is that its performance can be improved almost linearly with the increasing corpus and rule base, and especially for translating between 2 Indian languages for informal usage; the good enough translation is useful since the languages have a common root and hence share a large number of words across the different languages.

We have studied an approach to domain adaptation for SMT that enriches standard phrase-based models with lexicalised word and phrase pair features to help the model select appropriate translations for the target domain (TED talks). In addition, we show how source-side sentence-level topics can be incorporated to make the features differentiate between more fine-grained topics within the target domain (topic adaptation). We compare tuning our sparse features on a development set versus on the entire in-domain corpus and introduce a new method of porting them to larger mixed-domain models. Experimental results show that our features improve performance over a MIRA baseline and that in some cases we can get additional improvements with topic features. We evaluate our methods on two language pairs, English-French and German-English, showing promising results.

We also studied various methods for computing word alignments using statistical or heuristic models. We consider the five alignment models presented in Brown, Della Pietra, Della Pietra, and Mercer (1993), the hidden Markov alignment model, smoothing techniques, and refinements. These statistical models are compared with two heuristic models based on the Dice coefficient. We present different methods for combining word alignments to perform a symmetrization of directed statistical alignment models. As evaluation criterion, we use the quality of the resulting Viterbi alignment compared to a manually produced reference alignment. We evaluate the models on the German-English Verbmobil task and the French-English Hansards task. We perform a detailedanalysis of various design decisions of our statistical alignment system and evaluate these on training corpora of various sizes. An important result is that refined alignment models with a first-order dependence and a fertility model yield significantly better results than simple heuristic models.

III. RELATED WORK

Initial research has been done to translate Indian languages, mostly focusing Hindi and Bengali. However, most of the focus is still rule-based approach because of the unavailability of parallel data to build SMT systems for these languages. An approach for English to Bangla MT has been proposed by Dasgupta et al., 2004 that uses syntactic transfer of English sentences to Bangla with optimal time complexity. In generation stage of the phrases they used a dictionary to identify subject, object and also other entities like person, number and generate target sentences. An example-based machine translation system for English to Bangla has been proposed by Naskar et al., 2006 which

identifies the phrases in the input through a shallow analysis, retrieves the target phrases using the example-based approach and finally combines the target phrases using some heuristics based on the phrase reordering rules from Bangla. They also discussed some syntactic issues between English and Bangla.

A method to analyze syntactically Bangla sentence has been proposed by Anwar et al., 2009 using context sensitive grammar rules which accepts almost all types of Bangla sentences including simple, complex and compound sentences and then interpret input Bangla sentence to English using a NLP conversion unit. The grammar rules employed in the system allow parsing five categories of sentences according to Bangla intonation. The system is based on analyzing an input sentence and converting into a structural representation (SR). Once an SR is created for a particular sentence it is then converted to corresponding English sentence by NLP conversion unit. For conversion, the NLP conversion utilizes the corpus. A phrase-based Statistical Machine Translation (SMT) system has been proposed by Islam et al., 2010 that translates English sentences to Bengali. They added a transliteration module to handle OOV words. A preposition handling module is also incorporated to deal with systematic grammatical differences between English and Bangla. To measure the performance of their system, they used BLEU, NIST and TER scores. Durrani et al., 2010 also made use of transliteration to aid translation between Hindi and Urdu which are closely related languages. Roy &Popowich, 2009 applied three reordering techniques namely lexicalized, manual and automatic reordering to the source and language in a Bangla English SMT system.

A Phrase based model approach to English-Hindi translation has been proposed by Singh et al., 2012 in which they discussed the simple implementation of default phrase-based model for SMT for English to Hindi and also give an overview of different Machine translation applications that are in use nowadays. Sharma et.al. 2011 presented English to Hindi SMT system using phrase-based model approach. They used human evaluation metrics as their evaluation measures. These evaluations cost higher than the already available automatic evaluation metrics. The methods based on tree to string mappings has been proposed by Yamada & Knight, 2001 in which source language sentences are first parsed and later operations on each node. (Eisner, 2003) presented issues of working with isomorphic trees and presented a new approach of non-isomorphic tree-to-tree mapping translation model using synchronous tree substitution grammar (STSG). Li et al., 2005 first gave idea of using maximum entropy model based on source language parse trees to get n-best syntactic reordering's of each sentence which was further extended to use of lattices. Further Bisazza& Federico 2010 explored lattice-based reordering techniques for Arabic-English; they used shallow syntax chunking of the source language to move clause-initial verbs up to the maximum of 6 chunks where each verb's placement is encoded as separate path in lattice. They also discussed the complete divergence between two languages. Vocabulary difference between Urdu and English has been discussed.

IV. METHODOLOGY

After going through literature survey, we propose a plan which contain following modules**:**

a. Web App Designing

- In this module we will design web app from where user can operate translation part.
- We will also put some Sindhi learning tutorials.

b. Language Translator Architecture

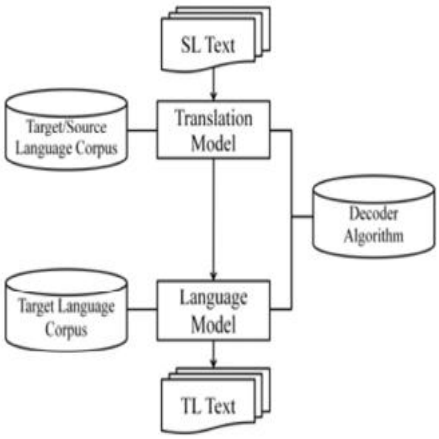

Figure 1: Example of language translator.

We present an analysis regarding the performance of the state-of-art Phrase-based Statistical Language Translation (SLT) on Indian Sindhi languages. We report baseline systems on Sindhi language pairs. The motivation of this study is to promote the development of SLT and linguistic resources for these language pairs, as the current state-of-the-art is quite bleak due to sparse data resources. The success of an SLT system is contingent on the availability of a large parallel corpus i.e. Dictionary. Such data is necessary to reliably estimate translation probabilities.

c. Dictionary Development

1. We will store vocabularies into dictionary.
2. We will store Grammar rules into dictionary.
3. We will store daily usable phrases and sentences into dictionary.
4. We will maintain grammar techniques into dictionary.

d. Phrase-based Model

In our experiment, we have used the Phrase-based SLT Models and evaluated their performance on the morphologically rich Indian languages. Phrase-based models are used to translate phrases of one or more words as atomic units. These models divide the input sentence into phrases, produce the target phrases and at the end reordering of these phrases is done.

e. Dataset

In Dataset We will load entire dictionary to dataset so as to apply vocabularies, grammar rules and techniques to translate from source to target phrase. Using Dataset will load everything fast, so translation time will be reduced.

CONCLUSION

In this research we employee Phrase-based model for training and used MERT for tuning our system. In this, we carried out a set of experiments by choosing the training, tuning and test sets from parallel corpus using the fivefold cross validation method to make up the fact that we had only a small amount of parallel data. We found that Indian Sindhi language got so much divergence when translating into English and that is why there is significant difference in obtained LT evaluation scores on seen corpus and on unseen test sets. In future, we will study SLT by applying other different approaches to develop good language models and also the training model whose more parallel corpus is available at the moment or may be available in nearer future.

FUTURE SCOPE

The developed SLT system takes the Indian language sentences as input and it generates corresponding closest translation in English. The translation of over 800 sentences was evaluatedusing automatic evaluation metric i.e. BLEU evaluation. BLEU scores may be concluded that the quality of translation is directly dependent on the scope and quality of parallel language corpora.

So, Future scope depends on corpus i.e. Vocabularies, Rules, Techniques stored into dictionary. Future Scope will be higher the volume of corpus into dictionary to increase the accuracy of translation.

REFERENCES

[1] NadeemJadoon Khan, Waqas Anwar & Nadir Durrani, "Machine Translation Approaches and Survey for Indian Languages",2017.

[2] ALPAC "Language and Machines: Computers in Translation and Linguistics".A report by the Automatic Language Processing Advisory Committee (Tech. Rep. No. Publication 1416), 2101 Constitution Avenue, Washington D.C., 20418 USA: National Academy of Sciences, National Research Council, 1966.

[3]Balajapally, P., Bandaru, P., Ganapathiraju, M., Balakrishnan, N., & Reddy, R., "Multilingual Book Reader: Transliteration, Word-to-Word Translation and Full-text Translation", 2006.

[4] Dwivedi, S. K., &Sukhadeve, P. P.,"Machine Translation System in Indian Perspectives",Journal of Computer Science, 6(10), 1111-1116, 2010.

[5] Antony P. J.,"Machine Translation Approaches and Survey for Indian Languages. Computational Linguistics and Chinese Language Processing", Vol. 18, No. 1, March 2013, pp. 47-78.

[6] Hasler, E., Haddow B., and Koehn, P.,"Sparse lexicalised features and topic adaptation for SMT",In Proceedings of the seventh International Workshop on Spoken Language Translation, pages 268–275, 2012.

[7] Och, F., "A systematic comparison of various statistical alignment models. Computational Linguistics", 29(1):19–5, 2003.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", In Proceedings of 40th Annual meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311–318.

[9] Bisazza, A. and Federico, M., "Chunk-based verb reordering in VSO sentences for Arabic-English statistical machine translation", In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, WMT '10,15-16 july 2010, pp. 235–243.

[10] J. Eisner, "Learning non-isomorphic tree mappings for machine translation", In Proceedings of the ACL Interactive Poster/Demonstration Sessions, 2003, 205–208.

[11] Dash, NiladriSekhar, Chaudhuri, BidyutBaran, "Why do we need to develop corpora in Indian languages?" A paper presented at SCALLA 2001 conference, Bangalore.

[12] Rao, Durgesh, "Machine Translation in India: A Brief Survey", SCALLA 2001 conference, Bangalore.

[13] Naskar, S., &Bandyopadhyay, S,"Use of Machine Translation in India: Current Status", In Proceedings of MT SUMMIT X; September 13-15, 2005, Phuket, Thailand.

[14] Bandyopadhyay S., "ANUBAAD - The Translator from English to Indian Languages" In proceedings of the VIIth State Science and Technology Congress, Calcutta. India, 2000, pp. 43-51.

Authors Profile

*Miss Pinky Gangwani*pursed Bachelor of Engineering in Computer Engineering from Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur in 2017and is currently a Master of Technology scholar from Rashtrasant Tukadoji Maharaj Nagpur University. Her main research work focuses on Statistical Language Translation (SLT),Phrase-based Translation,Natural Language Processing(NLP),Sindhi Phrase Word By Word.

