Review: English to Khasi Translation System

Bethsheba J. Rapthap
Computer Science & Engineering and IT
Assam Don Bosco University, Azara, Near Airport Road, Guwahati, Assam-781017, India

Pranab Das, PhD
Computer Science & Engineering and IT
Assam Don Bosco University,Azara, Near Airport Road,Guwahati, Assam-781017, India

ABSTRACT
This paper focuses on developing a translator which is translated from the English language to the Khasi language. Khasi is an Austroasiatic language spoken primarily in the North East India; Meghalaya state by the Khasi people. The paper uses the Rule-Based approach to perform translation. It takes as input sentences in the source language which is in English and outputs a translation of the sentences in the target language which is in Khasi.

Keywords
Rule-Based approach

1. INTRODUCTION
Languages work as an integrated and interconnected system. Basically, it connects the world with the human brain. The conceptualization of the world by perception and naming results in building the basic units of language, the words. Languages become the input into the brain from which the output of communication is made. Human beings more or less look at the world with similar experience and make abstractions about the world by conceptualization so that it is possible for them to exchange their ideas without difficulty. Language works at different levels of sounds, words and sentences. As sounds do not have any meaning on its own, hence, human beings depend on words as basic units of communication. We visualize, conceptualize and name things based on our conceptual precepts. While doing so, we make classifications of the objects of the world depending on their physical and telic properties.

India is a country with many languages. Khasi is one of the many languages spoken in the North East, India. Khasi is an Austroasiatic language spoken primarily in Meghalaya state by the Khasi people.

Every language has its defined structure with its own agreed upon rules. The complexity and singularity of the structure can be difficult for translation. Sentences in English has an order that comes in the form of subject, verb and object. For example, "he plays football." But not every language shares this structure.

Translation is the act of converting of text from one language to another. Hence this paper focuses on building a translator who will help users to be able to translate sentences in the English Language which is the source language to sentences in Khasi language which is the target language.

2. RELATED WORKS
2.1 Techniques
A method for Machine Translation (MT) of two Kurdish dialects is suggested, i.e., Kurmanji and Sorani. Both of these dialects are considered to be mutually unintelligible. The research used bi-dialectal dictionaries Machine Translation and the main goal is to show that the lack of a parallel corpora is not an obstacle in Machine Translation between the two dialects i.e., Kurmanji and Sorani[1].

The paper focuses on the translation of English sentences to Mauritian Creole language and from Mauritian Creole language to English sentences. Rule-Based Machine Translation approach is used to perform the translation. Input sentences are taken in the source language either in English or Mauritian Creole and it translates the sentences in the target language which is the output, either in English or Mauritian Creole[2,4,11]. Rule-Based approach is used in translating from English to Filipino text. It incorporates learning the Rule-Based on the analysis of a bilingual corpus in an attempt to eliminate the need for a linguist. The learning algorithm used in the paper is based on seeded version space learning algorithm. An open-source Rule-based Machine Translation system is developed for Scots. Scots is a low-resourced minor language closely related to English and spoken in Scotland and Ireland. For the development of dictionaries, by concentrating on translation for assimilation (gist comprehension) from Scots to English, it is proposed that the designed is to be use within the Apertium platform which will be sufficient to translate the improvement of non-scots speakers to understand the language.

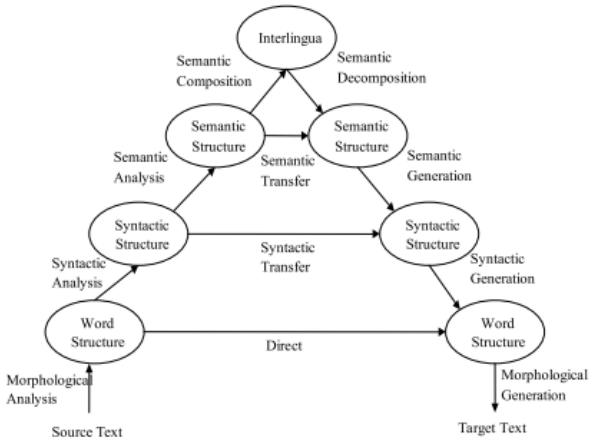

Figure 1: The Vauquois Triangle for MT

An English to Filipino Machine Translation system is develop. It is a bidirectional Machine Translation. The approach is to use a hybrid of Rule-Based and Example-Based paradigms by learning rules from examples. The paper uses a hybrid approach based on fixed rules and Transformation-Based Learning (TBL) method. The purpose of the paper is to transfer the English word orders into the Vietnamese. The learning process is train on the annotated bilingual corpus namely EVC: English-Vietnamese Corpus which will be word-aligned, phrase-aligned and POS-tagged. The development of Spanish to English translation task using a hybrid approach system is combine in a Phrase-Based Statistical Machine Translation (PBSMT) system. A bilingual information is obtain from a parallel corpora and a bilingual information from the Spanish to English language pair is also obtain in the Apertium rule-based machine translation (RMBT) platform[3,6,12].

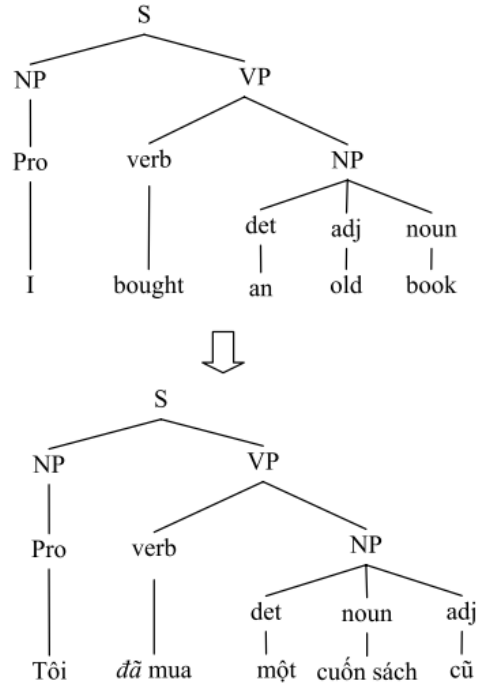

Figure 2: Transfer of English vs. Vietnamese trees

The paper mentioned in [5] successfully applied the hand-crafted morphological (de-)segmentation of Turkish, syntax-based pre-ordering of English in English-Turkish and post-ordering of English in Turkish-English. The paper performs desegmentation using SMT and propose a simple yet efficient modification of post- ordering.

A Statistical Machine Translation (SMT) system using MOSES is describe for translating a bidirectional English to Vietnamese translation system. Moses is an open-source toolkit for Statistical Machine Translation. The quality of a SMT system depends on the bilingual sentence alignment data called parallel corpus[7,8,10]. SMT and their two approaches along with their application to a task of English-to-Latvian translation are made. The outputs of the two Statistical Machine Translation is compared along with two different approaches to MT. Automatic evaluation metrics is used for reporting results. The development of the continued system of English to Irish domain tailored SMT system (Tapadóir) which is used currently by an in-house translation team of the Irish government department. The paper uses the Automated Post-Editing Module to improve the post-editing job of the translators.

A system for English to Yorùbá text translation process is describe. Natural Language Tool Kits (NLTKs) is used for testing and designing the rewrite rules. The formulated model was analyzed using Parse tree and Automata Theory-Based techniques. For the software design, Unified Modeling Language (UML) is used. The Python programming language and PyQt4 tools were used to implement the model[9].

A method for English-Hungarian translation is presented in the paper [13]. The translation is done by applying the reordering rules before the translation process and by creating morpheme-based and factor models.

2.2 Data Sets

For dictionary development, the paper [1] used web data, mainly websites of Kurdish media and universities in Iraqi Kurdistan region, for data collection. Regarding the genre, the paper selected the texts that were about art, literature, sport, and education.

To perform the translation, the paper [2,4,11] used the Rule-Based machine system which uses a bilingual dictionary to store the English language and the Mauritian Creole language. The paper has used the Diksioner Morisien dictionary to build the bilingual dictionary in the database. The goal is to provide an accurate translation. So that the quality of the translation is not affected, the paper makes sure that a series of tests were carried out whenever new rules were added. For the translation phase, the paper used a bilingual lexicon which contains the English word, its Filipino equivalent, and POS tag is use. Morphological analyser handles the inflections in the corpus and the input sentences. For the development of a dictionary, a mono and bilingual Scots dictionaries are constructed using lexical items which were gathered from a variety of resources across several domains. The development of the bilingual and Scots dictionaries contain around 3,000 individual lemmas (plus 5,758 personal names).

In the first phase, the combined Rule-Based and Example-Based paradigm which uses the Learning Module Architecture for the development of the reverse translation. The reverse translation is impossible since the problem lies in its dependency over the annotated grammar which is currently unavailable for Filipino. The second phase addresses this limitation by using information taken from English and Filipino POS Taggers. The second phase makes use of POS Taggers instead of using a parser[3,6,12]. The English-Vietnamese is built using the CADASA corpus. The corpus is organized into 24 text files. The English-Vietnamese sentences contain 8553 pairs. For the training phase 8053 pairs were used. The rest, i.e., 500 pairs, were saved for testing. A Phrase-Based statistical MT system is obtain. The phrase table is enrich with bilingual phrase pairs matching transfer rules and dictionary entries from the Apertium shallow-transfer Rule-Based MT platform.

The paper mentioned in [5] focus on post-processing: desegmentation of Turkish for English to Turkish and post-ordering of English words for Turkish to English. The paper employs additional SMT decoders to solve both tasks, which results in two-stage translation. It tried both Rule-Based or supervised and unsupervised approaches for morphological segmentation and English to Turkish reordering.

The paper mentioned in reference [7,8,10] shows the English-Vietnamese parallel corpus. Sentence pairs contains 882,000

and Vietnamese monolingual corpus from diversified resources contains 11,758,352 sentences. JRC Acquis parallel corpus of about 5.4M running tokens in the Latvian part and of about 6.7M tokens in the English part of the corpus is used in the paper mentioned in reference. Development set contained 500 sentences randomly extracted from the bilingual corpus; test corpus size was 1000 lines. An Automated Post Editing (APE) module is used. This module is used for checking the spelling issues that challenged the Tapadóir system. Using Automated Post Editing (APE), language style can be given more importance and can dedicate more time using translators. Using APE, it can help in improving the translator user-experience and any negative impact of repetitive grammatical or orthographic errors can be avoided. Thus, it helps in creating a more enjoyable user experience.

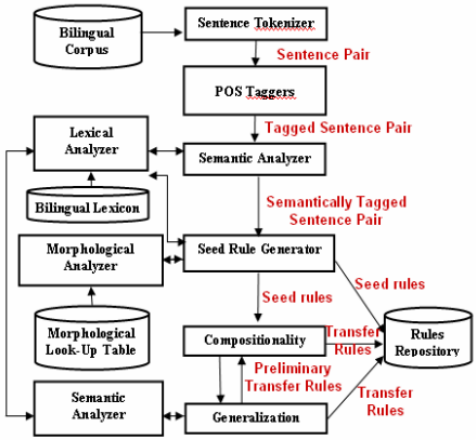

Figure 3: Learning Phase Architecture

The paper mentioned in reference [9] used English, and its equivalence in Yorùbá was collected using the home domain terminologies and lexical corpus construction techniques. The English to Yorùbá translation process was modeled using phrase structure grammar and rewrite rules.

The Hunglish corpus is used for the datasets. This corpus contains parallel texts from different domains such as literature and magazines, legal texts and movie subtitles. The corpus used for training the system consists of 1,026,836 parallel sentences with 14,553,765 words on the English side and 12,079,557 on the Hungarian side[13].

2.3 Evaluation Method

Several parameters were used in the paper mentioned in [1] such as fidelity or accuracy, intelligibility or clarity, and style. A combination of qualitative/quantitative approach to the evaluation process is followed. In the adaptation of human raters, the translated texts were given to several speakers whose main dialects are not the same as the original text. Also, some speakers will be chosen who have learned one of these dialects as a second language, and they do not have any familiarity with the other dialect. Quantitatively, the method evaluates the understandability degree of the translated texts using this parameter. The paper also conducts a short interview with the human raters after they rated the text to qualitatively assess the result. The results showed that the translated texts are understandable and it is 71% for Kurmanji and 79% for Sorani. They are rated as slightly understandable in 29% cases for Kurmanji and 21% for Sorani.

Table 1: Understandability of the IMT output - The table shows that 82% of the human raters, rated the output of IMT to be quite understandable

Understandability	Sorani to Kurmanji	Kurmanji to Sorani
Not Understandable	0%	0%
Slightly Understandable	29%	21%
Understandable	63%	71%
Completely Understandable	8%	8%

The goal is to provide the most accurate translation, therefore, whenever new rules were added, a series of tests were carried out to make sure that it does not affect the quality of translation[2,4,11]. To be accepted as a permanent rule, verification should be made beforehand that the system is able to translate 80% of the sentences. The sentences are translated based on the translating rule. With consideration, this is translated to the errors which may occur in the encoding of both the Morphological Analyzer and Lexicon. Hence, the sentence pairs were able to generate correct and accurate and the accuracy is 74%. A variety of evaluation methods were used, including a close test undertaken by human volunteers. For the evaluation, texts are taken at random from the Scottish Corpus of Texts &Speech.

In the first phase, the paper mentioned in [3,6,12] described the English to Filipino Translator which focused on the development of an English to Filipino Translator. It combines Rule-Based and Example-Based paradigm. In the second phase, it describes English Filipino Bilingual Translator, The original intent was to use the same architecture as the first phase in the development of the Filipino to English MT system. The learning process which is trained on the annotated bilingual corpus namely EVC: English-Vietnamese Corpus, has been word-aligned, phrase-aligned and POS-tagged automatically. For the transfer module in the English-Vietnamese transfer-based Machine Translation system, this transfer result is being used. A hybrid system for the Spanish to English language pair built by the given strategy was submitted. The initial phrase table was built from all the parallel corpora. The language model from the Europarl and the News Crawl monolingual English corpora. The weights of the different feature functions were optimised using minimum error rate training. The system also built a baseline PBSMT system trained on the same corpora and a reduced version of the system whose phrase table was enriched only with dictionary entries.

A hand-crafted morphological (desegmentation of Turkish) is successfully applied in the paper [5], syntax-based pre-ordering of English in English-Turkish and post-ordering of English in Turkish-English. The system performs de-segmentation using SMT and propose a simple yet efficient modification of post-ordering.

For translation evaluation, the paper uses BLEU (Bi-Lingual Evaluation Understudy). BLEU is a method of automatic Machine Translation evaluation. The translated output of the test set is compared with different manually translated references of the same set. According to the obtained BLEU scores, the proposed system performs better than the Google translator and the Microsoft Bing translator in both English to Vietnamese and Vietnamese to English translation[7,8,10]. On the first step, the paper is presented to the judge the PB-u and NB for every non-repetitive test line, which was then instructed to decide that the two translations were of equal

quality, or that one translation was better than the other. The difference in a total number of errors is negligible, however, a subjective evaluation of the system's output shows that the translation generated by the N-gram system is more understandable than the phrase-based one. The experiments to evaluate the addition of the APE module is described. The training data used to train and test the Tapadóir system is summarized. The paper then highlights the BLEU score changes following the introduction of the APE module. Some of the improvements which is introduced by the APE from a post-editing perspective is also discussed in the paper. The results may not always be reflected in an increase in BLEU scores.

Questionnaire design, detailing Expert and Experimental subject respondents (ESRs) profiles, and questionnaire administration are presented in the evaluation process[9,13]. Results of automated evaluation using a single reference BLEU metrics is presented. Translations were also generated by each system using human evaluation and by applying the ranking scheme to officially rank systems.

3. CONCLUSION

Translation is more than just translating of words or sentences from one language to another. A deep understanding of both grammar and culture is required. The rules of a language as well as the habits of the people who speak it needs to be focused by the translators. Hence this paper helps users to understand the Khasi language by translating sentences from the English language to sentences in the Khasi language.

4. ACKNOWLEDGEMENT

It gives me immense pleasure to express my deepest sense of gratitude and sincere thanks to my highly respected and esteemed guide and Head of the Department Dr. Pranab Das, for his valuable guidance, encouragement and help for completing this work. His useful suggestions for this whole work and cooperative behavior are sincerely acknowledged.

I would like to express my sincere thanks to the Dept. of Computer Science, School of Technology, Assam Don Bosco University, for giving me this opportunity to undertake this project. I am also grateful to my teachers for their constant support and guidance.

At the end I would like to express my sincere thanks to all my friends and others who helped us directly or indirectly.

5. REFERENCES

[1] "Kurdish Interdialect Machine Translation Hossein Hassani"; University of Kurdistan Hewler; Sarajevo; 2017.

[2] "English to Creole and Creole to English Rule-Based Machine Translation System"; Sameerchand Pudaruth, Lesh Sookun, Arvind Kumar Ruchpaul; Computer Science and Engineering University of Mauritius Mauritius; 2013.

[3] "Learning Translation Rules for a Bidirectional English-Filipino Machine Translator"; Michelle Wendy Tan, Bryan Anthony Hong, Daniel Liwanag Alcantara, Amiel Perez, and Lawrence Tan; Taft Avenue, Manila, Philippines; 2006.

[4] "Learning Translation Rules from Bilingual English – Filipino Corpus"; Michelle Wendy Tan, Raymond Joseph Ang, Natasja Gail Bautista, Ya Rong Cai, Bianca Tanlo; De La Salle University-Manila, Philippines; 2005.

[5] "Yandex School of Data Analysis approach to English-Turkish translation at WMT16 News Translation Task"; Anton Dvorkovic, Sergey Gubanov and Irina Galinskaya; Yandex School of Data Analysis, Timura Frunze St., Moscow, Russia; 2016.

[6] "A Hybrid Approach to Word Order Transfer in the English-to-Vietnamese Machine Translation"; Dien Dinh, Nguyen Luu Thuy Ngan, Do Xuan Quang, Van Chi Nam; IT Faculty, Vietnam National University of HCM City; HCM City,Vietnam; 2003.

[7] "Building a Bidirectional English-Vietnamese Statistical Machine Translation System by Using MOSES"; Nguyen Quang Phuoc, Yingxiu Quan, Cheol-Young Ock; Department of Electrical/Electronic and Computer Engineering, University of Ulsan, Ulsan, Korea; NAVER LABS, NAVER Corporation, Gyeonggi-do, Korea; 2016.

[8] "Towards improving English-Latvian translation: a system comparison and a new rescoring feature"; Maxim Khalilov, Jose A. R. Fonollosa, Inguna Skadina, Edgars Bralitis, Lauma Pretkalnina; Institute for Logic, Language and Computation, University van Amsterdam, Amsterdam, The Netherlands; Centre de Recerca TALP, University Politecnica de Catalunya, Barcelona, Spain; Institute of Mathematics and Computer Science, University of Latvia, Riga, Latvia. 2009.

[9] "Development of an English to Yorùbá Machine Translator"; Safiriyu I. Eludiora; Obafemi Awolowo University, Department of Computer Science & Engineering, Ile-Ife, Nigeria; Odetunji A. Odejobi, Obafemi Awolowo University, Department of Computer Science & Engineering, Ile-Ife, Nigeria; 2008

[10] "Tapadoir: English to Irish Machine Translation with Automatic Post-Editing"; Meghan Dowling Teresa Lynn Yvette Graham John Judge ADAPT Centre, Dublin City University, Dublin, Ireland; 2016.

[11] "A Rule-based Shallow-transfer Machine Translation System for Scots and English", Gavin Abercrombie; 2001

[12] "The Universitat d'Alacant hybrid machine translation system for WMT"; Victor M. Sanchez-Cartagena, Felipe Sanchez-Martinez, Juan Antonio Perez-Ortiz; Transducens Research Group; Department de Llenguatges i Sistemes Informatics; Universitat d'Alacant, Alacant, Spain; 2011.

[13] "An English to Hungarian Morpheme based Statistical Machine Translation System with Reordering Rules"; Laszlo J. Laki, Attila Novak, Borbala Siklosi, MTA-PPKE Language Technology Research Group, Pazmany Peter Catholic University, Faculty of Information Technology,Budapest,Hungary;2013.