ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 1 |January 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 7.032||ISI Value: 1.188

# CREATING AN EFFICIENT PARALLEL CORPUS FOR BANGLA-ENGLISH STATISTICAL MACHINE TRANSLATION

**A. A. Tanveer Hossain**[1]
[1]Department of Computer Science and Engineering,
United International University

**Foysal Ahmed Bhuiyan**[2]
[2] Department of Computer Science and Engineering,
United International University

**Md. Adnanul Islam**[3]
[3]Department of Computer Science and Engineering,
United International University

## ABSTRACT

***Parallel corpora are essential resources for multifarious linguistic analyses. In this article, we collect the data mostly from OPUS corpus, and preprocess the data for Bengali→English machine translation manually. We use MOSES (an SMT toolkit) for translating the Bengali texts. In this article, we elaborately discuss about collecting and formatting data, and an efficient translation mechanism from Bengali to English. To this context, we also evaluate the performance or translation accuracy in terms of BLEU, METEOR, and TER score.***

**KEYWORDS:** ***parallel corpus; machine translation; statistical machine translation; natural language processing***

## INTRODUCTION

A parallel corpus is a collection of text and each text is translated to another language. Here in this case only two languages are involved - English and Bengali. Using this collection, we will make a web based translator. So parallel corpus translation can be either unidirectional, bidirectional and multi directional. Our Translation is unidirectional, i.e., from Bengali text translated into English text. Collecting large-scale parallel corpus is needed huge attention.

A plethora of people are not well at interacting with English. Therefore, our study aims to motivate the Bengali monolingual speakers to learn the English language very easily, as learning English language is quite important at present. Moreover, an efficient machine translator relies heavily on the availability of necessary parallel corpora, which is lacking for Bangla-English language pair [11, 13]. Considering these aspects, the major objectives of this study are as follows:

1. We collecting a large amount of Bangla-English parallel sentences.
2. We perform necessary preprocessing by cleaning and formatting the collected data.
3. Next, we train MOSES using our created dataset.
4. Using MOSES, we translate Bengali texts to English.
5. Finally, we test the output for measuring the translation performance using our created parallel corpus.

## LITERATURE REVIEW

Parallel corpora proved to be extremely advantageous for cross linguistic research and translation in recent years. Although Bangla language is spoken by more than 210 million people all over the world as a first or second language, English has become the prominent language in today's world [12, 14]. To understand English, we aim to create an efficient Bangla→English translator. In existing literature, several approaches have been used to create bilingual or monolingual corpora. We analyze several such techniques with higher accuracy, which are summarized in Table 1.

**Table 1: Summary of several existing studies**

| Research paper | Year | Key contribution(s) |
|---|---|---|
| Resnik et al. [1] | 2004 | This study describes techniques for mining the Web to extract the parallel text it contains. STRAND is an architecture for structural translation recognition and acquiring natural data. |
| Mumin et al. [2] | 2012 | This study introduces an English-Bengali sentence-aligned parallel |

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 1 |January 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 7.032||ISI Value: 1.188

| | | |
|---|---|---|
| | | corpus consisting of more than 200,000 sentences. |
| Khosla et al. [3] | 2018 | This work presents a survey of the existing methods of building a parallel Corpus. Moreover, it discusses sentence alignment approach, web mining approach, manual approach, and machine translation approach. |
| Hasan et al. [4] | 2019 | This work describes how the performance of the models differ based on the data and modeling techniques. It also compares the results obtained from Google's machine translation system. |
| Rahman et al. [5] | 2010 | It shows how to use the root word to translate into English Sentence from Bangla sentence. It deals first to find out the root word from the database by defining the parts of speech of that sentence. Then, it detects the proper grammatical structure of that sentence for the target language. |
| Garg et al. [6] | 2011 | This study uses grapheme based method to model the transliteration problem by achieving an accuracy of 93.22%. |
| Akan et al. [7] | 2020 | This research focuses and investigates the complex issues of transliteration and translation from the Bangla into English, and tries to find suitable solutions on the basis of a need-based analysis and argument. |
| Mumin et al. [8] | 2001 | This paper describes an implementation process of a Machine Translator system between Bangla and English. Linguistic Knowledge (LK) architecture is preferred to implement this architecture. ln this system, Lexical Functional Grammar (LFG) framework is used . |
| Uddin et al. [9] | 2005 | This study utilizes some complex Bangla sentence types, and tries to solve them by creating some new parameters. |
| Hasan et al. [10] | 2019 | This work shows different NMT Bidirectional Long Short algorithms Term Memory (LSTM) and Transformer based NMT, to translate the Bangla to English language. |

In summary, none of these studies consider a plausibly large amount of data for Bangla→English translation. Moreover, average accuracy or performance level is quite low for Bangla→English translation using the existing approaches.

## METHODOLOGY

Our proposed methodology undergoes several major steps as mentioned below:

- Obtain the raw data
- Extract and align the document
- Prepare the corpus for statistical machine translation (SMT) systems

Figure 1 reflects the proposed methodology of our article.

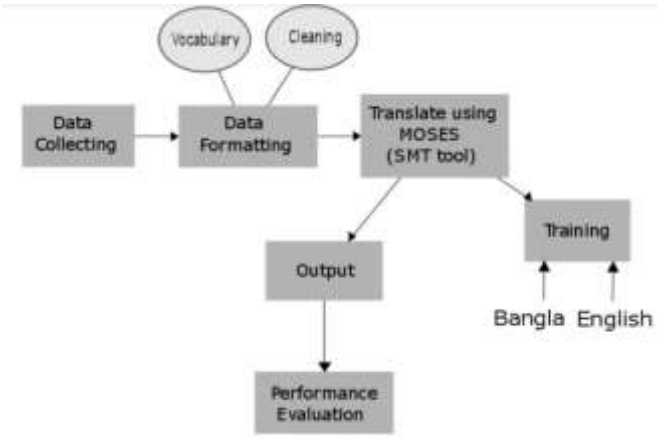


**Fig. 1: Proposed methodology**

### Data Cleaning and Formatting

The process of data cleaning is to detect and correct or remove inaccurate data from the dataset. Here, we have cleaned inaccurate or invalid characters from the corpus that are not in either Bangla or English form. For formatting, we manually align several Bangla-English sentence pairs, and perform proper translations wherever incorrect translations are found.

### Translation Using MOSES (SMT Toolkit)

MOSES is an implementation of the statistical or data-driven approach to machine translation. This is one of the superior approach as an SMT tool in the field at the moment. It is employed by the online translation systems established by the likes of Google and Microsoft. Here, we use standard external tools for some of the language modeling and processing tasks such as GIZA++, SIRLM.

### Training

The training dataset consists of input vector pairs and parallel output vector. We use both Bangla and English data, and the output is Bangla to English translation. In MOSES, the training process takes in the source and target data and uses occurrences of words and segments to conclude translation connects between the two languages.

### Training Pipeline and Decoder

The two main components in MOSES are - 1. The training pipeline and 2. The decoder. The training pipeline is a collection of tools which is written in Perl, with some in C++. It takes the raw data, which turn it into a machine translation model. Generally, the data need to be prepared and explicitly cleaned before starting training, tokenizing the data and converting tokens to a standard case. The decoder is a single C++ application. It mainly takes the trained machine translation model and source data to be translated as inputs, and translates the source data into the target language.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 1 |January 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 7.032||ISI Value: 1.188

### Performance Evaluation

Next, we perform the performance evaluation of Bangla→English machine translator using our dataset. We evaluate the translation performance using three standard performance metrics – BLEU, METEOR, and TER [15].
BLEU is one of the most popular automated and inexpensive metrics, which evaluate the quality of the machine translated text. BLEU also proposes an algorithm, called 'corpus bleu' to calculate the BLEU score for multiple sentences, for example - a paragraph or an article. The TER score measures the amount of editing that a translator would have to perform to change a translation so it exactly matches a reference translation. By repeating this analysis on a large number of sample translations, it is possible to estimate the post-editing effort required for a project. In TER a higher score is a sign of more post-editing effort and so the lower the score the better, as this indicates less post-editing is more required. METEOR can now automatically learn a version for a new target language using only the parallel data used for MT system development. The METEOR automatic evaluation metric scores machine translation hypotheses by aligning them to one or more reference translations.

## IMPLEMENTATION AND RESULTS

### Environment Setup

MOSES is a SMT tool which offers two types of translation models: one is phrase-based and another is tree-based. MOSES features organized translation models that enable the unification linguistic at the word level. Moses then allows new source language text to be decoded using trained statistical models to produce automatic translations in the target language. Figure 2 and Figure 3 show how we install MOSES in our machine using Linux operating system.


**Fig. 2: Importing MOSES from GIT**


**Fig. 3: Installing MOSES with Boost**

GIZA++ (Figure 4) is an SMT toolkit which used to train IBM Models 1-5 and an HMM word alignment model. It uses these models to compute Viterbi Alignments for SMT. It is an implementation of IBM model and it treats word alignment. It improved incomprehension calculation for models IBM-1, IBM-2 and HMM. It implements pegging, implemented a series of heuristics in order to make pegging properly efficient.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 1 |January 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 7.032||ISI Value: 1.188

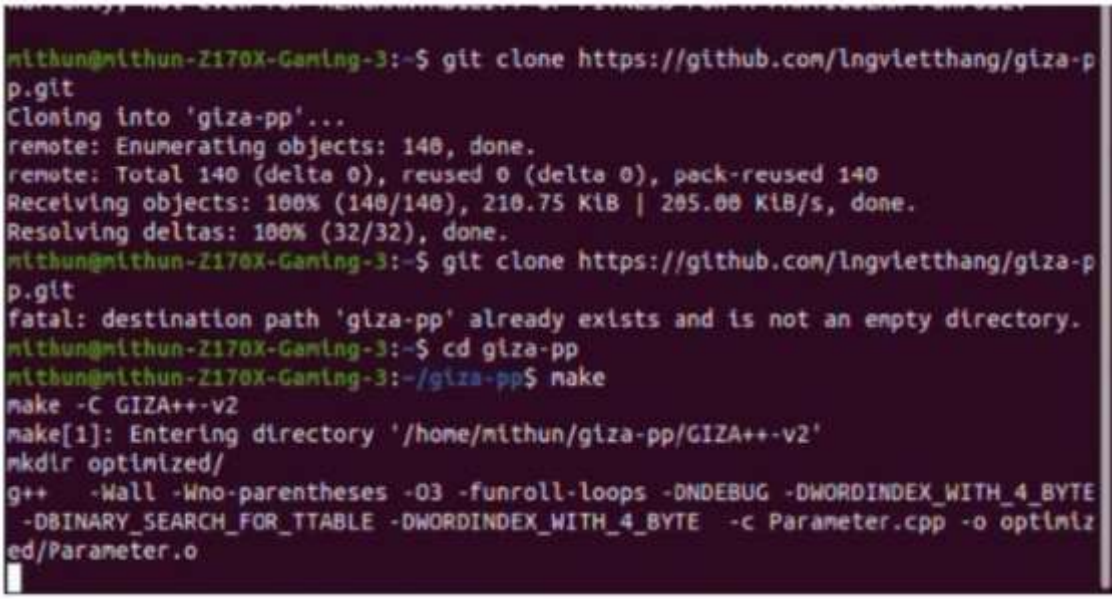


**Fig. 4: Installing GIZA++**

SRILM (Figure 5) is a toolkit which support Moses SMT system hierarchical phrase based SMT system. The toolkit can be downloaded and used free of charge. The main features of SRILM are to Generate the n-gram count file from the corpus and train the language model from the n-gram count file and calculate the test data incomprehension using the trained language model. To train the model properly, we have to make source and target languages word alignment using GIZA++, phrase extraction and scoring. It creates lexicalized reordering tables and configuration file of MOSES. To train 3.5 hundred thousand data this took about 5 hours using 4 cores on a powerful desktop (Intel i5-6500, 16GB RAM, AMD Radeon RX570 GPU). After finishing the training, it creates a 'moses.ini' file in the designated directory. We can use the model specified by this ini file to translate.

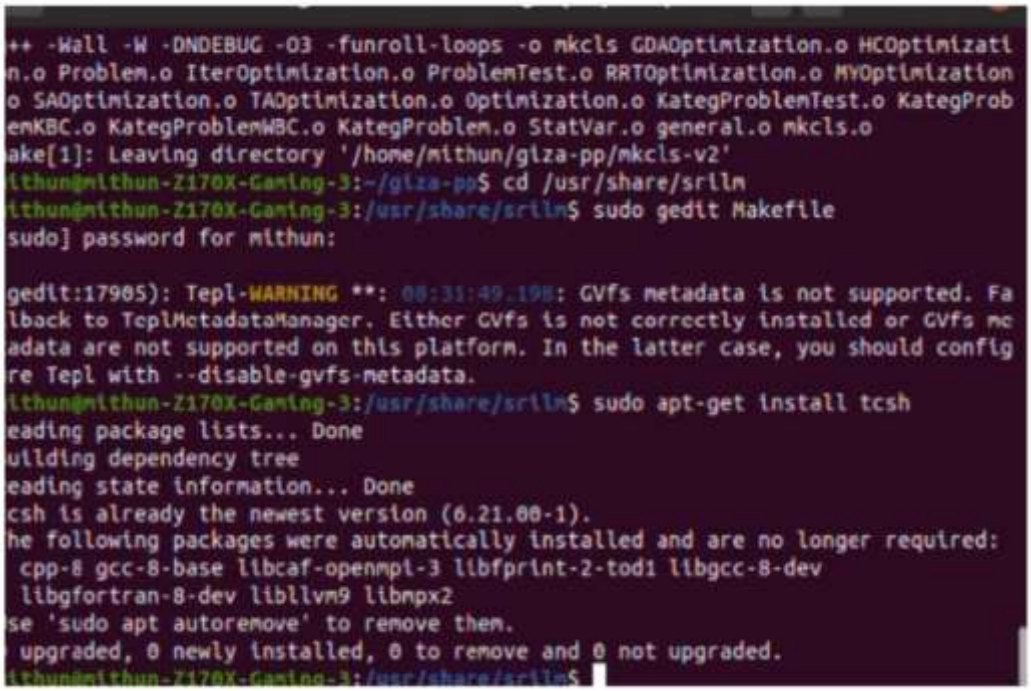


**Fig. 5: Installing SRILM**

## RESULTS AND DISCUSSIONS

We divide our corpus into training, development, and testing data. Note that, training and testing are mutually exclusive of each other. For decoding process (i.e., generating translations), we use a subset of our Bangla-English parallel corpus, which contains 1,000 Bangla source sentences. After testing theses source sentences with MOSES, the generated translations are reasonably close to the reference sentences (Figure 7).

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 1 |January 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 7.032||ISI Value: 1.188

This process of generating translation is reflected in Figure 6 and Figure 7.

```
mithun@mithun-Z170X-Gaming-3:~/Desktop/Work/model$ ~/mosesdecoder/bin/moses -f m
oses.ini.1 < /home/mithun/Desktop/Work/model/in > out
Defined parameters (per moses.ini or switch):
        config: moses.ini.1
        distortion-limit: 6
        feature: UnknownWordPenalty WordPenalty PhrasePenalty PhraseDictionaryMe
mory name=TranslationModel0 num-features=4 path=/home/mithun/Desktop/Work/model/
phrase-table.1 input-factor=0 output-factor=0 LexicalReordering name=LexicalReor
dering0 num-features=6 type=wbe-msd-bidirectional-fe-allff input-factor=0 output
-factor=0 path=/home/mithun/Desktop/Work/model/reordering-table.1.wbe-msd-bidire
ctional-fe.gz Distortion KENLM name=LM0 factor=0 path=/home/mithun/Desktop/Work/
lm/toy.binlm.1 order=5
        input-factors: 0
        mapping: 0 T 0
        weight: UnknownWordPenalty0= 1 WordPenalty0= -1 PhrasePenalty0= 0.2 Tran
slationModel0= 0.2 0.2 0.2 0.2 LexicalReordering0= 0.3 0.3 0.3 0.3 0.3 0.3 Disto
rtion0= 0.3 LM0= 0.5
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 0 end: 0
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=PhrasePenalty
FeatureFunction: PhrasePenalty0 start: 2 end: 2
line=PhraseDictionaryMemory name=TranslationModel0 num-features=4 path=/home/mit
```

**Fig. 6: Decoding by MOSES**

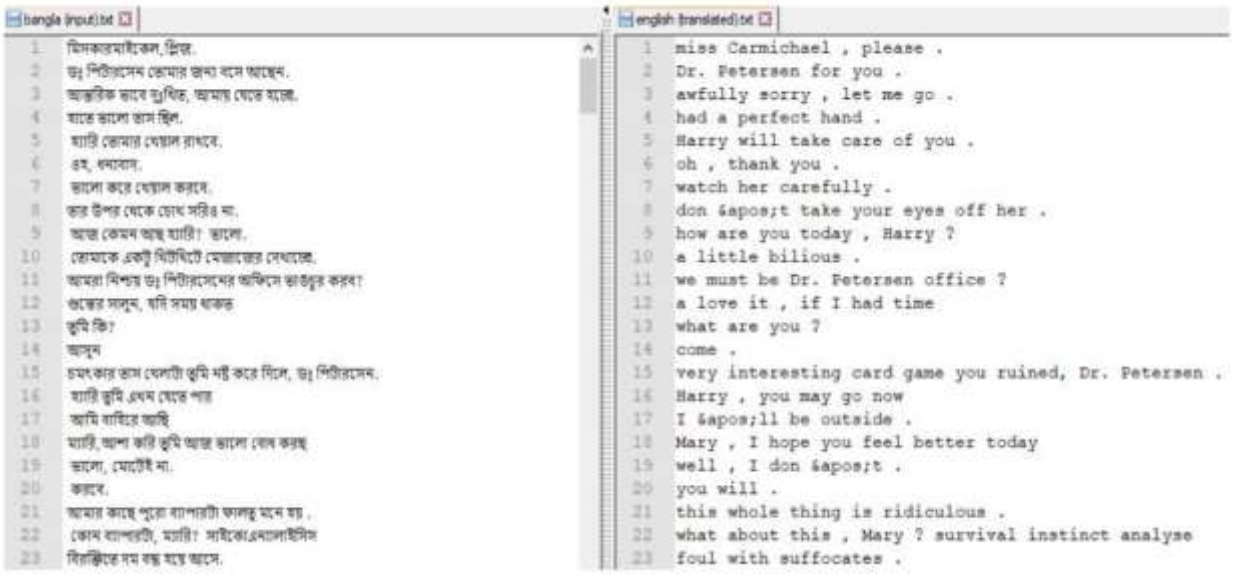


**Fig. 7: Snippet of our generated Bangla to English translations**

Next, we show the evaluation of our translations in terms of three different performance scores – BLEU, METEOR, and TER in Table 2. In Table 2, we find that our achieved BLEU, METEOR, and TER scores are 19.2, 27.7, and 83.2 respectively. Note that higher BLEU and METEOR scores and lower TER score indicate better performance.

**Table 2: Performance scores of Bangla→English translation using our dataset**

| Performance score | SOTA (Baseline) | MOSES with our dataset | Improvement over SOTA |
|---|---|---|---|
| BLEU | 18.5 | **19.2** | 4% |
| METEOR | 25.6 | **27.7** | 8% |
| TER | 86.6 | **83.2** | 4% |

These scores create a benchmark performance for state-of-the-art (SOTA) Bangla→English machine translation approaches. This result indicates the efficacy of our created novel parallel corpus for Bangla-English

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7 | Issue: 1 |January 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 7.032||ISI Value: 1.188

translation, which can be a useful resource for the researchers in future.

## CONCLUSION

Translators play a very important role in different real life applications [11]. Now-a-days, most of the global organizations thrive for higher accuracy in translating from one language to the desired language. Hence, our goal is to make such a tool that offers higher translation performance and also can achieve proper market value. However, such translation tools require rigorous training with more than a million parallel data of the highest quality. Thus, time and resource become immediate limitations of machine translation systems. Moreover, corpus-based translators (e.g., MOSES) cannot realise the highly sensitive rules of grammatically rich languages such as Bangla, since they do not explicitly deal with rules. Nevertheless, adding rules is a tough and time consuming task. Besides, ensuring quality data for training a machine translator seems to be a never-ending process. Therefore, the future work of this study includes enhancing the quality of our employed parallel Bangla-English corpus. Furthermore, our aim is to develop this tool for the low-resource Asian languages in order to translate them into English language. Apart from this, we plan to develop groundbreaking algorithm(s) to effectively realise the sensitive rules from quality parallel corpora.

## REFERENCES

1. *Noah. Resnik, Ps Smith. The web as a parallel corpus. computational linguistics. 2002.*
2. *Abu Shoeb Md Selim Mohammad Iqbal Muhammed Mumin, Mohammad Awal. Supara: A balanced english-bengali parallel corpus. 2012.*
3. *Sonal Khosla. A survey report on the existing methods of building a parallel corpus. International Journal of Advanced Research in Computer Science., 2018.*
4. *Firoj Chowdhury Shammur Khan Naira. Hasan, Md. Arid Alam. The web as a parallel corpus. computational linguistics. 2019.*
5. *Sangita Ph. D. M. Huda Mohammad. Rahman, Md Poddar. Open morphological machine translation: Bangla to english. 2010.*
6. *Vishal. Garg, Kamal Goyal. Development of a punjabi to english transliteration system. International Journal of Computer Science and Communication., 2011.*
7. *Faruquzzaman. Akan. Transliteration and translation from bangla into english: A problem solving approach. 2020.*
8. *Mohammad. Mumin. An implementation of machine translation between bangla and english. 2001.*
9. *Mahbub Hasan Muhammad. Uddin, Mohammad Murshed. A parametric approach to bangla to english statistical machine translation for complex bangla sentences. 2005.*
10. *Firoj Chowdhury Shammur Khan Naira. Hasan, Md. Arid Alam. Neural machine translation for the bangla-english language pair. 2009.*
11. *Md. Adnanul Islam and A. B. M. Alim Al Islam, "*Polygot: Going Beyond Database Driven And Syntax-based Translation*", in Proceedings of the 7th Annual Symposium on Computing for Development, ACM, 2016.*
12. *Md. Adnanul Islam, Md. Saidul Hoque Anik, and A. B. M. Alim Al Islam, "*Polygot: An Approach Towards Reliable Translation By Name Identification And Memory Optimization Using Semantic Analysis*", in Proceedings of the 4th International Conference on Networking, Systems and Security (NsysS), IEEE, 2017.*
13. *Md. Saidul Hoque Anik, Md. Adnanul Islam, and A. B. M. Alim Al Islam, "*An Approach Towards Multilingual Translation By Semantic-Based Verb Identification And Root Word Analysis*", in Proceedings of the 5th International Conference on Networking, Systems and Security (NSysS), IEEE, 2018.*
14. *S. Sakiba, M. Shuvo, J. Mela, B. Saha, N. Sultana, and M. Islam,* "A Memory-Efficient Tool for Bengali Parts of Speech Tagging". *In: Hemanth D., Vadivu G., Sangeetha M., Balas V. (eds) Artificial Intelligence Techniques for Advanced Computing Applications. Lecture Notes in Networks and Systems, vol. 130, Springer, 2020.*
15. *H. Hasan, M. Islam, M. Hasan, A. Hasan, S. Rumman, and N. shakib,* "A Spell-checker Integrated Machine Learning Based Solution for Speech to Text Conversion", *in Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology (ICSSIT), IEEE, 2020.*