Building a Sinhala-English Parallel Corpus for Neural Machine Translation Based on Exam Questions

MRM Rilfi[1#], UGYM Gunawansha[2], KAC Prasandika[2] and KGA Chandrani[2]

[1]*Inoovalab Technologies, Sri Lanka*
[2]*University of Moratuwa, Sri Lanka*

[#]rilfi@inoovalab.org

Abstract— In any neural machine translation between two natural languages, parallel corpus is a compulsory part of the training process. The most crucial step in an MT system is to develop an effective method for gathering parallel corpus. The construction of a parallel corpus, on the other hand, necessitates substantial knowledge of both languages and is a time-consuming procedure. Due to these limits, digitizing documents becomes extremely challenging, lowering the quality of machine translation systems. This research offers a method for producing an English to Sinhala parallel corpus that is both faster and more efficient, while requiring less human intervention. This system generates a parallel corpus for language pair using the following steps: scanning the exam question papers using a special type of scanner, Image optimization for Optical Character Recognition (OCR), text extraction from images and converting unstructured text into structured form as parallel corpus.

***Keywords:* parallel corpus, image optimization, text extraction, neural machine translation**

I. INTRODUCTION

English is widely used in both formal and informal communication nowadays. The mother tongue, on the other hand, continues to have a profound influence. The ability to think creatively and acquire global knowledge requires the use of one's mother tongue. According to Hettige et al., (2016) the Sinhala language is spoken by around 16 million people in Sri Lanka, and 80% of Sinhala speakers struggle to read and write English.

This is a language barrier that impedes the acquisition of world information. The most realistic answer to this challenge would be a computer-based machine translation system from English to other languages (B.Hettige & Karunananda, 2010)

Current machine transformation research necessitates large parallel corpus SMT and NMT systems based on probability models created utilizing parallel corpus components (Premjith, et al., 2019). However, one of the fundamental disadvantages of SMT systems is the lack of a parallel corpus. As a result, designing a machine translation engine for low-resource language pairs is extremely difficult (Doru, et al., 2018). The challenge of creating such corpses is regarded as enormous due to the considerable number of human resources required and the time required to produce a corpse with such many words. (Premjith, et al., 2016)

A parallel corpus is a collection of text translated into another language or saved in a machine-readable format from one or more languages. At the sentence or word level, a parallel corpse can be arranged (Hameed, et al., 2016) The quality of the parallel corporation used for system training is critical to the relevant system's success. With such a vast amount of data, it is simple to choose a domain-specific parallel company, and the full training data may be processed in a short length of time (Doru, et al., 2018). The length of the sentences is also a significant factor in determining the translation's quality. It should not be too short or too long in length. This is because deep learning architecture is incapable of absorbing the extensive dependencies seen in phrases. Word phrases have a vital role in domain-specific machine translation, according to Teenage et al. They investigate effective strategies of employing phrases to increase NMT performance under the low Resource Languages heading (Tennage, et al., 2017). Because these languages' morphological richness lies on opposite extremes of the spectrum, translation can be improved by including linguistic features in the phrase (Premjith, et al., 2019).

II. LITERATURE REVIEW

A. Neural Machine Translation

Neural Machine Translation is becoming the current state of modern machine translation technology. Although NMT has been successful for resource languages, its relevance in less resource settings are still controversial (Tennage, et al., 2017). However, Neural machine translation is a recently proposed approach to machine translation. Unlike traditional statistical machine translation, it aims to build a single neural network that can be tuned together to maximize performance (Bahdanau, et al., 2015).

B. History of machine translation

In the seventeenth century, the dream of the translation of natural languages by machine became the reality in the late twentieth. (Hutchins, 1995). The first automated translation systems were developed in 1933 by Georges Artzoni and Peter Troyansky. Warren Weaver briefly touches on early perception research as an effective way to transform machinery in 1949 and the first set of proposals for computer-based machine translation was put forward in 1949 by Warren Weaver. Sixty years later, neural networks have made considerable progress in other areas but are still unable to convince translators. The use of MT accelerated in the 1990s. This increase is most noticeable in commercial companies, public service, and multinational companies. Where translations are mass-produced, primarily technical documentation (Hutchins, 1995). The first active neural language models powered by repetitive neural networks appeared in 2011(Scao , 2020).

C. Existing Work

Machine translation systems are divided into four categories: human-assisted translation, rule-based translation, statistical translation, and example-based translation. Each of these approaches to machine translation has its own set of benefits and drawbacks (Hettige, et al., 2011).Agent-based Multi-agent system for language processing applications such as English to Sinhala (Hettige, et al., 2016)Another group, Wijerathna et al and De Silva et al, used a statistical methodology to attempt English-to-Sinhala machine translation. The authors collect bilingual corpora and examine parallel corpora; many corpora have 100,000 parallel sentences per language pair. NMT is discussed in many studies, with a focus on zero shot neural machine approaches.

(Hettige, et al.,2017). Another group attempted English-to-Sinhala machine translation by using a statistical methodology. Wijerathna et al. and De Silva et al. presented simple rule-based translators. Hettige et al. have provided a theoretical-based method for English-Sinhala machine translation based on the Sinhala concept of Varanagema (conjugation). Varanagema is a Sinhala language theory that deals with nouns, verbs, and prepositions, among other things. Tennage et al. developed a domain specific NMT system for both Sinhala and Tamil, the official languages of Sri Lanka. The translation of official government documents was focused on that research. They used the NMT architecture suggested by Bahnao et al. and Cho et al. For all experiments. From the Sinhala language point of view only a few research have been done for machine translation. Vitanage's English to Sinhala translator and Silva and other Sinhala to English translator are some of the prototype projects for the weather forecasting domain. There have been some attempts at machine translation from Sinhala to Tamil and from Japanese to Sinhala machine translation (Hettige &Karunananda, 2010)

Kumar and Sarawagi investigated calibration of state-of-the-art models that can be improved by a parametric model, resulting in a slight increase in BLEU score. However, neither work investigates the relation between label smoothing during training and calibration. (Kumar & Sarawagi, 2019). B. Hettige and Karunananda(2010) found that, in the post-editing phase of the translation process, using human intervention could solve the problems caused by the lack of dictionary and semantic information for high quality translation. (Hettige, B. et al.,2017) Minh-Thang Luong et al. demonstrated the effectiveness of both approaches on the WMT translation tasks between English and German in both global and local directions. (Minh-Thang Luong, et al., 2015) Thilakshi Fonseka et al. introduced an effective NMT system along with Byte Pair Encoding (BPE) for the English-Sinhala language pair focusing on the Sri Lankan official government documents. It addressed the OOV problem and the data sparsity

issue when translating to a more morphologically rich language. (Thilakshi Fonseka, et al.,2020).
Toshiaki Nakazawa et al. developed the "Kyoto-U system" that attended the IWSLT06 Japanese-English machine translation task. The system consists of two modules, the alignment module for the parallel sentences and the translation module for obtaining and integrating appropriate translation examples. (Toshiaki Nakazawa, et al.,2006). Francisco Guzman et al. introduced the FLORES evaluation datasets for Nepali English and Sinhala– English, based on sentences translated from Wikipedia. Their experiments represented current state-of-the-art approaches that work poorly with these new benchmarks, with semi-supervised, especially multilingual neural methods, all other models they have considered. (Francisco Guzman, 2019).

III. PROPOSED SOLUTION

The proposed method is creating a parallel corpus in an effective and efficient manner by using ScanJet Pro 2000 Scanner and Google Optical Character Recognition (OCR). For turning photos into text format, the approach makes use of Google OCR's efficiency. A scanner can scan a wide range of documents, and this scanner can scan any document quickly and without causing any damage to the original (as we do with typical scanners). It can scan papers in the following sizes: A3, A4, B5, B4, and so on. For scanned images, Google OCR produces satisfactory results. The OCR technique can be used to automate several photos with only minor changes. The parallel sentences must be aligned, which necessitates text normalization effort. Then we normalize raw data set to database. Finally, we are generating a parallel corpus.

A. Image Pre-processing

Because the amount of picture and video data generated and consumed daily is expanding, the need for better and more efficient image modification techniques is also expanding. Recent advancements in image processing have sparked renewed interest in neural networks. In the field of image processing, a type of neural network known as Convolutional Neural Networks is particularly intriguing since it offers a novel method to comprehending image data.

For this project, many documents were digitized from numerous sources that included both English and Sinhala materials. Even though identifying and collecting documents with such paper pairs was a big undertaking, countless exam papers were checked, and many parallel papers were picked out.

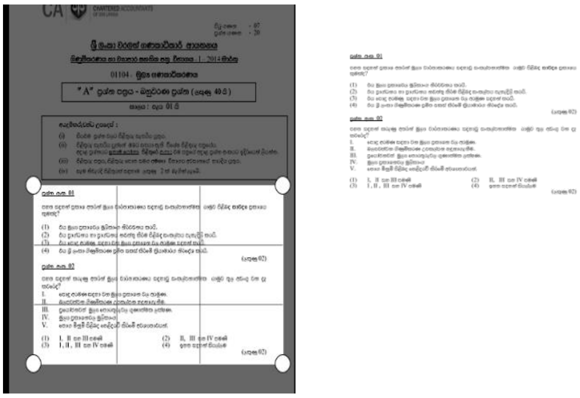


Figure 16.Image Cropping

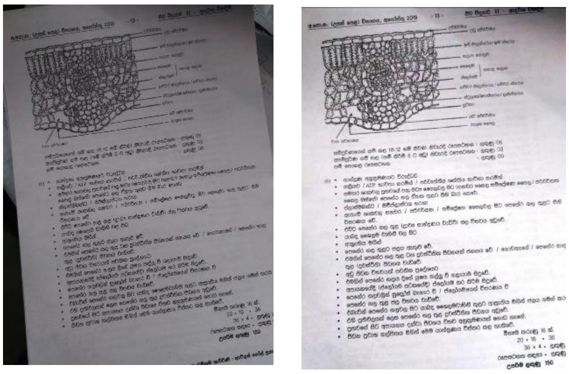


Figure 18. Original Image (Left) and Transformed Image (right)

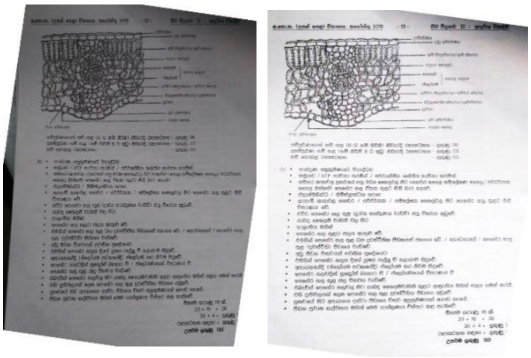


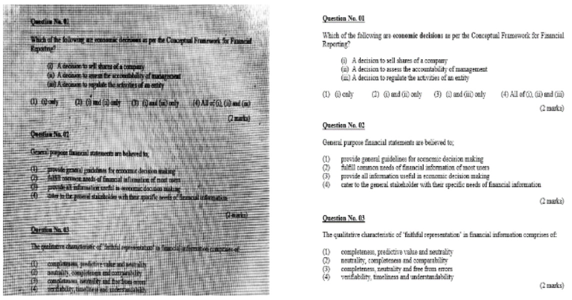
Figure 17.Color Correction

A few preparations must be completed before scanning can begin. Documents must be put at a reasonable height because the scanner light is very intense. The positioning of documents is another key factor to consider. We will not acquire a good OCR of the appropriate document if the alignment of the document is not correct and if the document is tilted or shaken during scanning. Even if we scan papers with the utmost care, some pre-processing is required because Google OCR produces excellent results for high-quality results reprocessing on the scanned pages, such as skew correction and cutting out of unwanted elements, is done before uploading to Google Cloud Storage. Brightness, contrast, and sharpness can all be tweaked to further improve image quality if necessary. This pre-processing stage can be completed with application. Parallel sentences were collected from sources such as GCE OL, AL, and available competitive examinations papers (both English and Sinhala medium). When scanning low-quality documents, such as old manuscripts, we get excessively noisy images, making excellent OCR impossible. This means that a text document with adequate alignment and lighting will produce accurate results using Google OCR.PDFs with many pages from these sources are split to obtain one PDF.PDF with two-column structures must be cut into two separate papers with the order. Cut the text in the bottom line and the heading of the paper. (Figure 1) Because the high-quality size must be lowered, scanned images can have a large size. The higher the quality of the source image, the better the OCR accuracy. There are certain factors that can be considered in the source image to measure the quality of the source image. Characters should stand out against the background: Character borders that are sharp, Characters / Words with a High Contrast Alignment: Proper letter, word, and line segmentation is ensured by good alignment. Image alignment and resolution (Figure 2,3). There is less noise. From an OCR standpoint, the qualities stated above improve the document quality. The quality of the source picture for OCR is determined by several parameters, including the presence or absence of noises/distortions, proper image and text alignment, image resolution, and local contrast. The standard recommended resolution for OCR is 300 DPI (Dots Per Inch) However, based on the font size used some OCR engines internally scales the original image. Image Binarization is the process of converting a colored image (RGB) into a black and white image.

Most of the OCR Engines does this process internally, Adaptive binarization works based on features of neighbouring pixels (i.e., local window) Sharp borders between characters will be helpful for character recognition. Image Despeckling is a common technique used in the OCR noise removal step which is an adaptive bilateral filtering technique (Figure 4). It removes noises from the scanned image while preserving edges and other complex areas from blurring. When applied incorrectly it may remove commas and apostrophe from the image by considering them as noise. Such changes should be made in this step. Image data collection diagram for corpus creation are given below. (Figure 5)

B. Extract Text from Images

Google Optical Character Recognition (OCR) is a software which works for over 248 languages in the world. It can detect many languages with over 90% of accuracy and it is simple and easy to use. This technology extracts scanned printed text, text from images and even handwriting. It uses the dependencies from Tesseract and released as a free software. This method utilizes the Google OCR as an API (application programming interfaces) and Google Cloud as a service for increasing the processing speed with the large amount of data. The optimization of Google OCR is at an elevated level and after conversion, information can be analysed with multiple different methods. Uploading an image or a pdf to Google Cloud Storage and using it as an input bucket yields Combustible uploading and OCR creation were automated because Google Cloud Storage only permits one document to be created at a time. Applying OCR for image/PDF diagram are given below. (Figure 6) Even if we upload high-quality documents, there is a chance that the OCR will contain errors. Manually correcting errors like spelling fixes, eliminating unwanted spaces and unwanted characters, and adding missing spaces and missing characters is simple. This method, in our experience, generates OCR of both English and Sinhala text documents with 98 percent accuracy, resulting in a high-quality input document. The final phase in constructing the parallel corpus is the most time-consuming because it necessitates a great deal of focus. Initially, we split one paper that includes many pages into one page PDF. This process will increase the amount of data in the corpus.



Question No. 01

Which of the following are **economic decisions** as per the Conceptual Framework for Financial Reporting?

(i) A decision to sell shares of a company
(ii) A decision to assess the accountability of management
(iii) A decision to regulate the activities of an entity

(1) (i) only (2) (i) and (ii) only (3) (i) and (iii) only (4) All of (i), (ii) and (iii)

(2 marks)

Question No. 02

General purpose financial statements are believed to;

(1) provide general guidelines for economic decision making
(2) fulfill common needs of financial information of most users
(3) provide all information useful in economic decision making
(4) cater to the general stakeholder with their specific needs of financial information

(2 marks)

Question No. 03

The qualitative characteristic of 'faithful representation' in financial information comprises of:

(1) completeness, predictive value and neutrality
(2) neutrality, completeness and comparability
(3) completeness, neutrality and free from errors
(4) verifiability, timeliness and understandability

(2 marks)

Figure 20.Original Image (Left), Despeckled Image (Right)

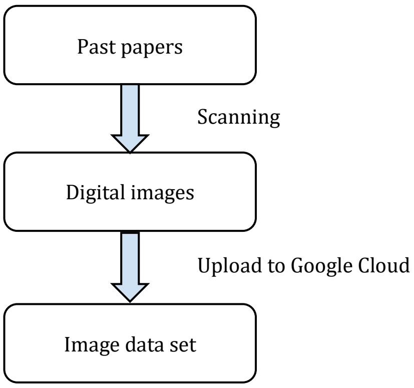


Figure 20.Diagram of image data collection

C. *Text Normalization*

After extracting sentences from the images, it is required to normalize the raw data. unstructured raw data contains image data, data of the footer section, instructions, logos, table contents, and other unwanted information. In normalization remove that unnecessary information. Manual work is required for parallel sentence alignment and text normalization. Regular expression is used to create text that conforms to the relevant question pattern. Text normalization diagram are given below. (Figure 7) Regular expression is one of the most useful tools in computer science. It is used in phonology, graphics, text analysis, information extraction, and speech recognition. Regular expressions are placed in the matching pair and describe the strings of characters. It is a pattern that matches certain strings and doesn't match others. Regular expression is a set of characters that define a pattern (Kaur, 2014).

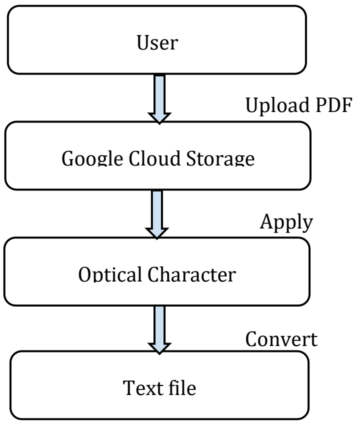


Figure 21.Diagram of applying OCR.

D. *Parallel Corpus Generation*

A parallel corpus is a corpus that contains a collection of original English writings as well as translations into the Sinhala language. Parallel corpora often contain data from only two languages. Building parallel corpus diagram are given below. (Figure 8) 'Comparable corpora,' which are intricately connected to parallel corpora, consist of texts from two or more languages that are similar in genre, topic, register, etc. but do not include the same content. This is a parallel bilingual corpus.

To create high quality custom engines, we need to have enough domain-specific open parallel corporations and both tools and methods to create this body (Dogru, 2018). One of the reasons for the low resource language scores lower in machine translation evaluation is that machine translation systems are usually trained with a small amount of data or low-quality data (Dogru, 2018)

E. *Text format*

To prepare for an exam, you must first determine not just the substance of the test, but also the types of questions that will be asked. Diverse types of questions necessitate different study methods. There are many distinct types of questions, as well as varied study and preparation tactics for each. Multiple choice, True/False, Matching, Short answers, Numerical, and Essay are some of the question types used to store text data taken from papers. Multiple-choice tests usually begin with a question or statement to which you must respond by choosing the best option from a list of options. True-false tests ask students to mark whether certain assertions are true or false. All aspects of a statement must be true for it to be true. True-false tests, in general, assess your understanding of facts. Preparing for true-false examinations requires the same general study abilities and best practices as studying for other types of tests. Students must respond to assertions or questions in essay questions. Short-answer questions or statements are like essay questions in that they require only a few words or sentences to answer. They assess basic information, which is usually factual. It is crucial to pay attention to the directive words in each item when answering short-answer questions.

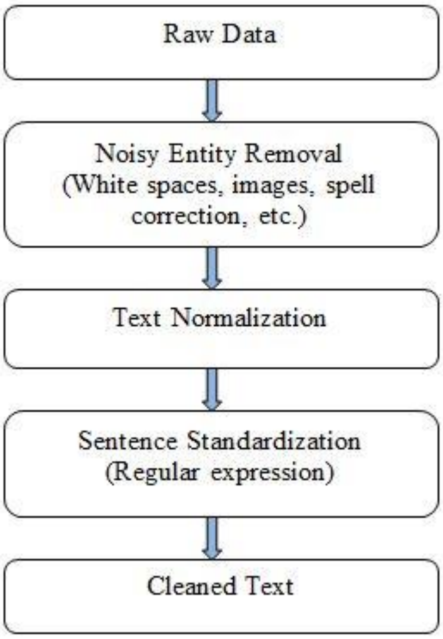

Raw Data
Noisy Entity Removal (White spaces, images, spell correction, etc.)
Text Normalization
Sentence Standardization (Regular expression)
Cleaned Text

Figure 7. Diagram of Text normalization

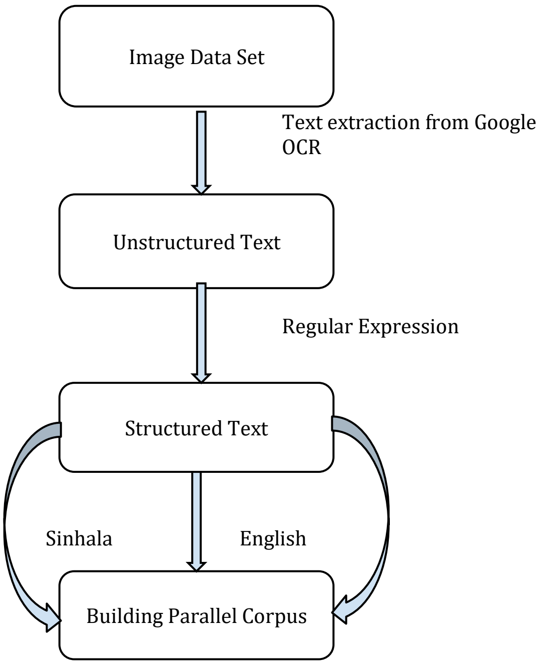


Figure 22.Diagram of building parallel corpus

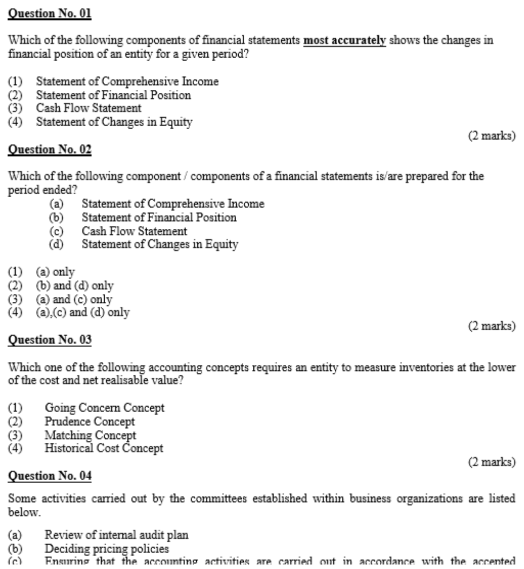
V. EXPERIMENTS AND RESULTS

First, scanned the exam question papers and prepared them as PDFs or images. A lot of documents collected from various resources which contain both English and Sinhala were scanned for this work. Many exams were searched and many parallel sentences in different languages. This was a massive task, but the results were incredibly good.

The educational realm is the focus of our model. As a result, government exam papers in both languages are originally collected. A particular data pre-processing approach is used to the papers. That is, text from all languages must be aligned so that corresponding segments, phrases, or paragraphs can be matched. The data set was first scanned. They must be scanned in a high-quality format that allows the text to be read clearly. The dataset is then saved on the cloud, making OCR easier to apply. Image removal, spell correction, and corrections in superscript subscript style are all required alterations to the text file created by OCR. The parallel corpora are required for the study. For pair of languages (English – Sinhala, English – Tamil) a parallel corpus is constructed. Models is included in the study, one for pair of languages. The data must be pre-processed before the model can be trained.

Because the models are based on the seq2seq architecture, an encoder-decoder architecture. It is made up of two LSTM (Long Short-Term Memory) networks [34], which are a form of RNN (Recurrent Neural Network). The encoder LSTM is one, while the decoder LSTM is the other.

Each RNN's input differs from the others. The encoder's input the images are shown in Figure 9.



Question No. 01

Which of the following components of financial statements **most accurately** shows the changes in financial position of an entity for a given period?

(1) Statement of Comprehensive Income
(2) Statement of Financial Position
(3) Cash Flow Statement
(4) Statement of Changes in Equity

(2 marks)

Question No. 02

Which of the following component / components of a financial statements is/are prepared for the period ended?

(a) Statement of Comprehensive Income
(b) Statement of Financial Position
(c) Cash Flow Statement
(d) Statement of Changes in Equity

(1) (a) only
(2) (b) and (d) only
(3) (a) and (c) only
(4) (a),(c) and (d) only

(2 marks)

Question No. 03

Which one of the following accounting concepts requires an entity to measure inventories at the lower of the cost and net realisable value?

(1) Going Concern Concept
(2) Prudence Concept
(3) Matching Concept
(4) Historical Cost Concept

(2 marks)

Question No. 04

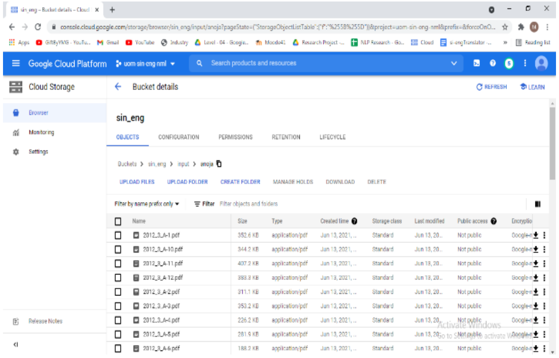
Some activities carried out by the committees established within business organizations are listed below.

(a) Review of internal audit plan
(b) Deciding pricing policies
(c) Ensuring that the accounting activities are carried out in accordance with the accepted

Figure 23. Image of scanned paper

The PDF or image prepared in this way is broken down into individual pages and stored in the Google Cloud Storage (Figure 10).

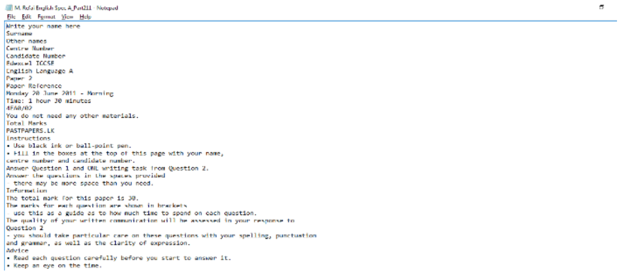
They are then taken out of the bucket and applied to it. Subsequent texts are rearranged using regular expressions and stored as parallel corpus. After extracting sentences from the images, it is required to normalize the raw data (Figure 11).



☐	Name	Size	Type	Created time	Storage class	Last modified	Public access	Encryptio
☐	2012_3_A-1.pdf	352.6 KB	application/pdf	Jun 13, 2021, ...	Standard	Jun 13, 20...	Not public	Google-n
☐	2012_3_A-10.pdf	344.2 KB	application/pdf	Jun 13, 2021, ...	Standard	Jun 13, 20...	Not public	Google-n
☐	2012_3_A-11.pdf	407.2 KB	application/pdf	Jun 13, 2021, ...	Standard	Jun 13, 20...	Not public	Google-n
☐	2012_3_A-12.pdf	383.3 KB	application/pdf	Jun 13, 2021, ...	Standard	Jun 13, 20...	Not public	Google-n
☐	2012_3_A-2.pdf	311.1 KB	application/pdf	Jun 13, 2021, ...	Standard	Jun 13, 20...	Not public	Google-n
☐	2012_3_A-3.pdf	353.2 KB	application/pdf	Jun 13, 2021, ...	Standard	Jun 13, 20...	Not public	Google-n
☐	2012_3_A-4.pdf	226.2 KB	application/pdf	Jun 13, 2021, ...	Standard	Jun 13, 20...	Not public	Google-n
☐	2012_3_A-5.pdf	281.9 KB	application/pdf	Jun 13, 2021, ...	Standard	Jun 13, 20...	Not public	Google-n
☐	2012_3_A-6.pdf	188.2 KB	application/pdf	Jun 13, 2021, ...	Standard	Jun 13, 20...	Not public	Google-n

Figure 24. Image of Google cloud storage.

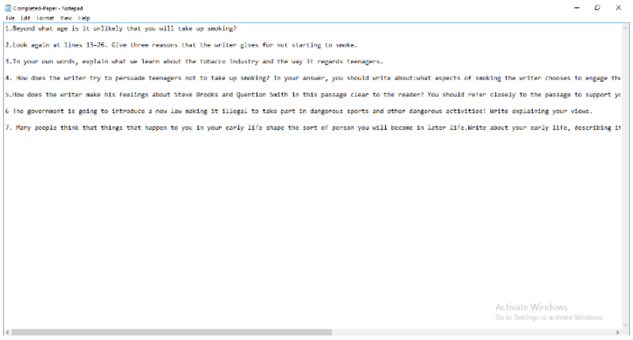
Normalization removes image data, data of the footer section, instructions, logos, table contents, and other unwanted information. Manual work is required for parallel sentence alignment and text normalization (Figure 12).



M. Rafai English Spec A_Part211 - Notepad
File Edit Format View Help

```
Write your name here
Surname
Other names
Centre Number
Candidate Number
Edexcel IGCSE
English Language A
Paper 2
Paper Reference
Monday 20 June 2011 - Morning
Time: 1 hour 30 minutes
4EA0/02
You do not need any other materials.
Total Marks
PASTPAPERS.LK
Instructions
• Use black ink or ball-point pen.
• Fill in the boxes at the top of this page with your name,
centre number and candidate number.
Answer Question 1 and ONE writing task from Question 2.
Answer the questions in the spaces provided
– there may be more space than you need.
Information
The total mark for this paper is 30.
The marks for each question are shown in brackets
– use this as a guide as to how much time to spend on each question.
The quality of your written communication will be assessed in your response to
Question 2
- you should take particular care on these questions with your spelling, punctuation
and grammar, as well as the clarity of expression.
Advice
• Read each question carefully before you start to answer it.
• Keep an eye on the time.
```

Figure 25. Image of before apply normalization to



Completed-Paper - Notepad
File Edit Format View Help
Activate Windows
Go to Settings to activate Windows.

```
1.Beyond what age is it unlikely that you will take up smoking?

2.Look again at lines 13-26. Give three reasons that the writer gives for not starting to smoke.

3.In your own words, explain what we learn about the tobacco industry and the way it regards teenagers.

4. How does the writer try to persuade teenagers not to take up smoking? In your answer, you should write about:what aspects of smoking the writer chooses to engage the

5.How does the writer make his feelings about Steve Brooks and Quention Smith in this passage clear to the reader? You should refer closely to the passage to support yo

6 The government is going to introduce a new law making it illegal to take part in dangerous sports and other dangerous activities! Write explaining your views.

7. Many people think that things that happen to you in your early life shape the sort of person you will become in later life.Write about your early life, describing it
```

Figure 26. Image of normalized paper

V. CONCLUSION AND FURTHER WORK

The purpose of this research was to develop Sinhala-English Parallel Corpus for Neural Machine Translation using an OCR Scanner and Google Optical character recognition (OCR). Exam question translation was focused on this research. Translation between English and Sinhala is difficult due to the different language structures in English and Sinhala. The lack of parallel corpus is a major drawback for SMT systems and therefore, designing a machine translation system for low resource languages is exceedingly difficult. However, a crucial factor to consider when preparing a parallel corpus is that the person who is preparing the corpus should have a good knowledge of both languages. However, using the proposed system could generate many parallel corpus without much knowledge of both languages in a brief period. The main task to be done further in this project is to develop an English-Sinhala Machine Translation System.

REFERENCES

B.Hettige & A. S. Karunananda, 2010. *Theoretical based approach to English to Sinhala.* Sri lanka, s.n.

Bahdanau, D., KyungHyun Cho & Yoshua Bengio, 2015. *NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE.* s.l., s.n.

Doğru, G., Adrià Martín-Mor & Anna Aguilar-Amat, 2018. *Parallel Corpora Preparation for Machine Translation of Low-Resource.* s.l., s.n.

Hameed, R. A. et al., 2016. Automatic Creation of a Sentence Aligned Sinhala-Tamil. *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing,* pp. 124-132.

Hutchins, W., 1995. *MACHINE TRANSLATION: A BRIEF HISTORY,* s.l.: s.n.

Premjith, B. et al., 2016. A Fast and Efficient Framework for Creating Parallel. *Indian Journal of Science and Technology,* pp. 1-7.

Premjith, B., M. Anand Kumar & K.P. Soman, 2019. Neural Machine Translation System for. pp. 387-398.

Tennage, P. et al., 2017. *Neural Machine Translation for Sinhala and Tamil Languages.* s.l., s.n.

Chen, H., Huang, S., Chiang, D., Chen, J., 2017. Improved Neural Machine Translation with a Synpax-Aware Encoder and Decoder.

de silva, D., Alahakoon, A., Udayangani, I., N P, V., Kolonnage, D., Perera, H., Thelijjagoda, S., 2009. Sinhala to English Language Translator. pp. 419–424. https://doi.org/10.1109/ICIAFS.2008.4783983

Hettige, B., 2011. A COMPUTATIONAL GRAMMAR OF SINHALA FOR ENGLISH-SINHALA MACHINE TRANSLATION. https://doi.org/10.13140/RG.2.1.2330.6968

Hettige, B., Karunananda, A., 2010. Theoretical based approach to English to Sinhala machine translation. pp. 380–385. https://doi.org/10.1109/ICIINFS.2009.5429832

Hettige, B., Karunananda, A., Rzevski, G., 2017. Phrase-level English to Sinhala machine translation with multi-agent approach. pp. 1–6. https://doi.org/10.1109/ICIINFS.2017.8300419

Hettige, B., Karunananda, A., Rzevski, G., 2016. A multi-agent solution for managing complexity in english to Sinhala machine translation. Int. J. Des. Nat. Ecodynamics 11, 88–96. https://doi.org/10.2495/DNE-V11-N2-88-96

MADHUBALA, D., RAO, N.K., ROOPA, D.Y.M., KALYANI, D., PRANAY, M., RAJU, C.K., n.d. A TOOL TO CONVERT AUDIO/TEXT TO SIGN LANGUAGE USING PYTHON LIBRARIES. Turk. J. Physiother. Rehabil. 32, 2.

Singh, M., Kumar, R., Chana, I., 2019. Encoding-Decoding Methods for Neural Machine Translation. pp. 1454–1459. https://doi.org/10.1109/ICICICT46008.2019.8993143

Tennage, P., Herath, A., Thilakarathne, M., Sandaruwan, P., Ranathunga, S., 2018. Transliteration and Byte Pair Encoding to Improve Tamil to Sinhala Neural Machine Translation. pp. 390–395. https://doi.org/10.1109/MERCon.2018.8421939

Wijerathna, L., Somaweera, W.L.S.L., Kaduruwana, S.L., Wijesinghe, Y., De Silva, D., Pulasinghe, K., Thelijjagoda, S., 2012. A Translator from Sinhala to English and English to Sinhala (SEES). pp. 14–18. https://doi.org/10.1109/ICTer.2012.6421408

Scao, T. L., 2020. *A brief history of machine translation paradigms.* [Online]Available at: https://medium.com/huggingface/a-brief-history-of-machine-translation-paradigms-d5c09d8a5b7e

Aviral Kumar, S. S., 2019. CALIBRATION OF ENCODER DECODER MODELS FOR. arXiv:1903.00802v1, p. 14.

B. Hettige, A. S. K., 2017. Developing Lexicon Databases for English to Sinhala Machine Translation. p. 6.

Minh-Thang Luong, H. P. ,. D. M., 2015. Effective Approaches to Attention-based Neural Machine Translation. p. 11.

Thilakshi Fonseka, R. N. R. P. U. T., 2020. English to Sinhala Neural Machine Translation. 978-1-7281-7689-5/20/$31.00.2020IEEE, p. 05.

Toshiaki Nakazawa, K. Y. D. K. S. K., 2006. Example-based Machine Translation based on Deeper NLP. International Workshop on Spoken Language Translation(IWSLT), p. 07.

Francisco Guzman, P.-J. C. F. M. O. F. J. P. G. L. P. K. V. C. R., 2019. The FLORES Evaluation Datasets for Low-Resource Machine Translation Nepali-English and Sinhala-English. p. 14.