A Comparison of Data Augmentation Techniques for Nguni Language Statistical and Neural Machine Translation models

Project Proposal

Fezeka Nzama
Department Of Computer Science
University Of Cape Town
Cape Town, South Africa
nzmfez001@myuct.ac.za

Deevesh Beegun
Department Of Computer Science
University Of Cape Town
Cape Town, South Africa
bgndee001@myuct.ac.za

CCS Concepts

• **Computing methodologies → Machine translation**.

Keywords

Neural Machine Translation, Statistical Machine Translation, Low Resource Language, Statistical Models

1 Project Description

The Nguni language group of South Africa, including isiZulu and isiXhosa, constitutes the most widely spoken language group in the country [1]. Despite this prevalence in everyday South Africa, Nguni language resources constitute less than 1% of content online, whereas more than 50% of internet content is written in English [2]. Given the importance of language in enabling trade, and exchanging knowledge, translation facilities are essential to the development of South African commerce and education. The use of human translators is ideal, but expensive and slow. The alternative of machine translation is more scalable, cheaper and faster.

Machine translation refers to the use of computer systems to translate some source text from a particular source language into a corresponding text in a target language [18]. This field of study makes use of statistical models, to learn the underlying characteristics of a source and target language pair and as such translate between the two.

The current state-of-the-art is neural machine translation (NMT), which makes use of a neural network architecture. This model has allowed for higher quality translations when compared to its predecessor, statistical machine translation, when applied in the high resource language setting [16, 22]. High resource languages are those languages for which there exist large amounts of parallel text (parallel corpora) for training purposes. However, these advancements have not been seen in the low and extremely low resource language setting, where SMT has generally been shown to outperform NMT [19, 27, 28] despite more recent research showing cases where NMT has outperformed SMT for short sentences, or when NMT models have been finely tuned [9, 27]. Nguni languages fall into this low resource language category, and as such research into machine translation for this language group is limited but provides an ideal scenario to investigate the use of SMT in a low resource setting, in contrast with NMT.

2 Problem Statement & Research Questions

The primary challenge faced by machine translation for Nguni languages is that of data sparsity. The models which have been proven to yield the best results, are undermined by the limited amount of existing parallel corpora for these languages. Whilst a number of data augmentation techniques have been suggested in past research [3, 6, 13, 15, 31], some of these techniques have not been applied in the Nguni language context and there is no existing performance comparison of these techniques for Nguni language machine translation. This presents a problem in that there is currently no established "best practice" data augmentation technique for Nguni language machine translation on which further research may be based. As such, this project aims to provide a comparison between baseline MT models trained on existing parallel corpora and MT models trained on data that is augmented using one of the following two techniques: 1) the creation of additional parallel corpora using monolingual corpora via back-translation and 2) the use multilingual parallel corpora as training data. The goal is to establish which, if either, of the aforementioned data augmentation techniques leads to the highest translation quality in the Nguni language context. To achieve this goal the following research questions are proposed:

(1) Does making use of parallel corpora augmented with synthetic parallel corpora from the back-translation of monolingual corpora as training data for MT models lead to higher BLEU scores when compared with the baseline MT models when trained on low resource Nguni languages?
(2) Does making use of multilingual parallel corpus as training data lead to higher BLEU scores for MT models, when compared to the baseline MT models for low resource Nguni languages?
(3) What combination of training data augmentation technique and machine translation models yields the highest BLUE scores for Nguni language translation?

3 Procedures and Methods

In this section we give a brief overview of the different approaches we are going to take to answer the research questions posed in section 2 and discuss the procedures and methods that we are going to use.

The machine translation models, namely, phrase-based SMT and Sequence to Sequence Transformers NMT, will be trained on several datasets with appropriate hyperparameter optimization to get the most optimal machine translation model. Publicly available datasets obtained from different sources will be used as training data. These datasets consist of monolingual and parallel data for the translation of English to isiXhosa and isiZulu. Data augmentation techniques such as back translation will be applied to convert the monolingual datasets into additional parallel data. In addition, by training the models using several parallel corpora consisting of isiZulu and isiXhosa, multilingual models will be created. Finally, the performance of the models will be evaluated by computing the BLEU scores for each model and appropriate quantitative comparisons will be made with the different models.

3.1 Datasets

Datasets	Number of tokens (million)
SADiLaR (parallel)	1.38
SADiLaR (monolingual)	2.42
Opus Corpus (parallel)	1.7
JW-300 corpus (parallel)	0.6
C4 multilingual datasets (monolingual)	60

Table 1: Training Data: English -> isiXhosa.

Datasets	Number of tokens (million)
SADiLaR (parallel)	1.0
C4 multilingual datasets (monolingual)	200

Table 2: Training Data: English -> isiZulu.

3.1.1 English -> isiXhosa For the translation of English to isiXhosa we will use datasets retrieved from The South African Center for Digital Language Resources (SADiLaR) [1] website which have been made available as a result of the Autshumato project. These datasets consist of a bilingual corpus, which have been aligned, containing translations from English to isiXhosa in two separate text files and a monolingual corpus [2] containing only isiXhosa sentences. These monolingual data will be converted to parallel data via back-translation. Datasets obtained from Opus Corpus [3] will also be used. These datasets which contain parallel corpora have been gathered from various places such as Medical charts (translated by the Medical Machine Translation project (MeMaT)), Bible, Cape Town bylaws, South African Constitution, Universal declaration of Human Rights, Mobile Xhosa, South African Navy from Stellenbosch, University of Cape Town Clinical, Crawled from Western Cape Government and Wiki titles. In addition, high-quality parallel data obtained from the JW-300 corpus [4] will be used as our main training data. A summary statistics for the datasets is provided in table 1.

3.1.2 English -> isiZulu Tokenised parallel text retrieved from The South African Center for Digital Language Resources (SADiLaR) [5] website will be used for training models to translate English to isiZulu. These parallel data have been extracted from various sources such as the Autshumato corpus, translated texts from Wikipedia, the Bible, the Book of Mormon, the Constitution of South Africa, the Universal Declaration of Human Rights and some sentences from the book "Beyond the He/Man". In addition, data obtained from the C4 multilingual dataset [6] containing monolingual data for isiZulu will be used as additional parallel data obtained via backtranslation. A summary statistics of the above datasets is provided in table 2.

3.1.3 Cleaning Data Since the performance of Machine Translation models relies heavily on the amount and quality of data available, data cleaning forms a crucial part in the development of an optimal model. Using poor data quality is more evident in Neural Machine Translation models where their performance degrades more than that of Statistical Machine [17]. This is due to NMT memorising bad examples more quickly [23].

A lot of the data that the models will be trained on are raw web crawl data and data obtained using back translation; hence, data cleaning will be of utmost importance. This process involves true casing, removing duplicate translations, and removing non-alphabetic characters [7]. Some corpora filters will be used to clean the data depending on the corpus that is being used. These filters are as follows[23]:

- Unique parallel sentence filter – removes duplicate source-target sentences.
- Equal source-target filter- removes identical source side and target side sentences.
- Multiple sources – one target and multiple targets – one source filters – removes sentence pairs where multiple source sentences are aligned with one target sentence and one source sentence is aligned with multiple target sentences.
- Non-alphabetical filters – remove sentences that contain non-alphabetical symbols.
- Repeating token filter – filters parallel corpora created from a monolingual corpus, using back-translation.
- Moses Scripts – calls Moses scripts for tokenising, cleaning and true casing.

3.2 Using Monolingual Data

The performance of a Machine Translation model depends on the amount of data the model has been trained on; consequently, the

[1] available at: https://repo.sadilar.org/handle/20.500.12185/525
[2] available at: https://repo.sadilar.org/handle/20.500.12185/524?show=full
[3] available at: https://opus.nlpl.eu/memat.php
[4] available at https://opus.nlpl.eu/XhosaNavy.php
[5] available at: https://repo.sadilar.org/handle/20.500.12185/489?show=full
[6] available on: https://github.com/allenai/allennlp/discussions/5265

more data we have, the better our model will perform. While phrase-based Statistical Machine Translation models have benefited by using monolingual data as training data, Neural Machine Translation uses only parallel data for training which are often sparse, especially for low resource languages [29]. On the other hand, in some cases, there is a substantial amount of monolingual data available for the target language.

Sennrich et al. [29] showed that monolingual training data can be treated as additional parallel training data which could improve the quality of translation models by mixing monolingual target sentences into the training set. They proposed two techniques to achieve this: the first one treats monolingual training examples as parallel examples with empty source side. The second technique that the author proposed is to pair monolingual training instances with a synthetic source sentence. This synthetic data is obtained via back-translation. It is performed by training a machine translation model that translates the target text into the source text. This model is then used to produce the synthetic data. In this project we will use the second technique since it resulted in a greater improvement in BLEU score compared to the first technique.

A combination of the monolingual corpus of the SADILaR and C4 multilingual datasets will be used to train two back-translation models one for isiZulu to English and another one for isiXhosa to English. The synthetic data produced by the back-translation model will then be combined with the monolingual dataset, thus creating additional parallel data. This data and the other parallel data from SADiLaR, Opus Corpus and JW 300 Corpus will be used to train the SMT and NMT models to translate from English to isiZulu and from English to isiXhosa.

3.3 Multilingual Machine Translation Model

The performance of Machine Translation models for low resource languages can be improved by training the models on a joint set of bilingual corpora with languages that have similar semantics [11]. This brings multilinguality which helps improve individual translations [14].

In this project we will use the method proposed by Ha et al. [14] for training the NMT models in a multilingual setting. This method uses a universal Encoder and Decoder, i.e using a single NMT system, thus does not require any modification to the architecture of the system. We will also perform a one to many translation i.e, from one source language to multiple target languages (English to isiXhosa and isiZulu) which Dong el al.[8] showed there is a significant achievement in translation quality over individually learned models when this method is used.

In the case of SMT, the translation model will be trained on a merged corpora of English-Xhosa and English-Zulu parallel corpora. Banerjee et. al. suggest the use of language model also trained on a merged corpora [5]. However, research into multilingual SMT is limited largely to pivot based multilingual examples of SMT. As such, the use of both a merged language model and target language specific language model will be explored during experimentation.

Both the SMT and NMT model will be trained on a merged corpus. This corpus will consist of parallel data from the SADiLaR, Opus corpus, and the JW 300 corpus for both English to isiXhosa and English to isiZulu translations. In the case of the NMT model a language token will be added in the input to differentiate from the different languages.

3.4 Subword Segmentation

Sennrich et al.[30] showed that models trained on sequences of subword units are more accurate than large vocabulary models. They also showed that new words that have not been seen during training time can also be generated.

Due to the agglutinative nature of Bantu languages, Byte pair encoding will be used as word segmentation to model open vocabulary translation in the machine translation models. We will use the method that Sennrich et al. used that does not require a back-off model for rare words, instead it encodes rare words via subword units [30].

3.5 Model Implementation

The phrase-based SMT model will be implemented using the Stanford Phrasal Toolkit. It is an open source, phrase-based MT toolkit written in Java. It provides multithreaded decoding and online tuning for learning feature-rich models for very large datasets [12]. Additionally, the KenLM language modelling tool will be used in conjunction with the Phrasal toolkit. This language modelling toolkit is faster and uses less memory than other leading language modelling toolkits.

The Sequence to Sequence Transformers NMT model will be implemented using the FairSeq Toolkit. It is a sequence modeling toolkit that allows to train custom models for translation, summarization, language modelling and other text generation tasks [25]. It is written in pytorch and supports distributed training across multiple GPUs and machines [25].

3.6 Model Validation and Evaluation

The models will be evaluated using k-fold cross-validation on parallel data obtained from the various sources and monolingual data converted to parallel data via back translation. The data will be split into three sets namely, train, test and validation. The train set will be used to train the models, the validation set will be used to evaluate the models during training and the best model will be evaluated on the test set.

The performance of the machine translation models will be measured using The Bilingual Evaluation Understudy (BLEU) method proposed by Papeinri et al [26]. BLEU is an automatic evaluation method; it is fast, inexpensive, provides an objective view and strongly correlates to human evaluation [10].

4 Ethical Professional and Legal Issues

Human evaluators will not be used in the testing process for this project. As such, no prior consent or ethical considerations are necessary.

Third-party machine translation toolkits will be used within this project. Thus, the sources of these toolkits will be acknowledged and the conditions of use followed stringently. Additionally, training data from third-party libraries will be utilised. The sources of these libraries will also be acknowledged.

5 Related Work

5.1 English to Nguni Translation

As previously mentioned, research into machine translation for low resource Nguni languages is limited. However, in a study published by Martinus and Abbot benchmarking NMT for Southern African languages, isiZulu was found to have the lowest Bleu scores when compared with 4 other South African languages, whilst Afrikaans had the highest [21]. A transformer model was trained on the Aushumato parallel corpora, achieving a BLEU score of 1.34 for isiZulu, with Afrikaans boasting a BLEU score of 20.60, the highest of all tested languages .This poor performance was attributed to a limited dataset for training, low quality data and the morphological complexity of the language. Griesel et. al suggest the use of statistical machine translation models with data pre-processing steps to render the Nguni languages more comparable to English for translation purposes [22]. The aim of this work is to filter out some of the differences presented by the morphological structure of Nguni languages in comparison to English.

5.2 Monolingual Data Augmentation Techniques

The use of synthetic parallel corpora created from monolingual corpora in the target languages has been suggested to yield improvements in translation quality. This is done by using a filtered back-translation approach, creating a synthetic source sentence paired with the target sentence [15, 20, 31]. Bertoldi and Federico suggest the use of source side monolingual corpora in creating synthetic parallel corpora for domain specific machine translation [6]. This method is expanded upon by the work of Wang and Zong, incorporating a domain dictionary into the model along with monolingual corpora [32].

5.3 Multilingual Data Augmentation Techniques

The idea of a universal machine translation has also been proposed, using a set of parallel corpora derived from parallel corpora with a single source language and different target languages [3, 13].Nyoni and Basset provide an argument for the use of multilingual learning in the low resource setting for English to isiZulu translation based on a English-to-isiXhosa and English-to-isiZulu learning model for NMT [24]. The model is trained on parallel texts from the Orpus corpus, Omniglot encyclopedia, Linguanaut Phrases Center and the Wild Coast Xhosa phrase book. This multilingual learning yields better results than the baseline or zero-shot learning approaches when trained with similar multilingual languages with a gain of 9.9 BLEU Score in comparison with the model baseline [24].

6 Anticipated Outcomes

In this section we describe the outcome of the project, the key features and major design challenges. We also discuss the impact that this project will have and how we evaluate its success.

6.1 System

This section describes the different models that will need to be implemented by the end of this project.

By the end of this project, the following models will need to be correctly implemented: Firstly, a baseline phrase-based Statistical Machine Translation (SMT) and a Sequence to Sequence Transformers Neural Machine Translation (NMT) model trained using parallel data obtained from the various sources will be implemented. Secondly, synthetic data generated from monolingual data via back translation will be used to train both the SMT and NMT models. The back translation will be performed using another translation model that translates the target sentences to the source sentences, i.e., from isiXhosa or isiZulu to English. Thirdly, the SMT and NMT models will be trained in a multilingual setting benefiting from the semantic similarities of both languages. All these models will need to be trained on high-quality data so that a fair comparison can be made between them. As such, the data will need to be preprocessed using byte-pair encoding and with appropriate data cleaning techniques. After the models have been trained they will be able to translate from English to isiXhosa, from English to isiZulu, and from English to isiXhosa and isiZulu in the case of the multilingual settings. Finally, the performance of all these models will need to be compared with each other and the one that gives the greatest improvement will be noted. This will help us conclude which technique helps achieve the greatest improvement in the translation of low resource South African languages.

6.2 Expected Impact

Translation plays a crucial role in the development and expansion of languages. It maintains cultural and linguistic diversity [4]. African languages, which are often sparse, have not gained much attention in the machine translation field. This project will help preserve these languages, specifically Nguni languages, by using machine translation to translate English to isiXhosa and isiZulu. It will also help traditional Bantu languages transition into the modern world where English is dominant. This will allow intercultural communication between African people and other people globally, thus moving towards an inclusive system.

We expect the multilingual models to result in a greater performance improvement than when using monolingual data to train the models. We also expect that the state of art Neural Machine Translation model will perform better than the Statistical Machine Translation models in low resource settings when these different data augmentation techniques are used.

6.3 Key success factors

For this project to be considered successful, it will need to satisfy various criteria. These criteria are as follows: high-quality parallel data will need to be generated via back translation of monolingual data. Maintaining the quality of the training data is crucial since it significantly influences the performance of machine translation models, especially for neural machine translation models. Furthermore, all models will need to be correctly implemented, with the NMT model outperforming the SMT models as anticipated. In addition, all evaluations of the models will need to result in a comparable BLEU score, and appropriate comparisons will need to be made between them so that the research questions can be sufficiently answered. Finally, proving a significant improvement in BLEU score for the different models compared to other research performed in the machine translation field for Nguni languages will result in considerable success for this project.

7 Project Plan

The following section outlines the plan that will be undertaken to complete this project.

7.1 Risks

Potential risks that may occur have been identified and outlined in the risk matrix attached in Appendix A. The probability and impact of each risk has been assigned a mark out of 10, where 0 indicates no probability or impact and 10 indicates a certainty or a devastating impact. The consequence of each risk coming to fruition as well as mitigation, monitoring and management strategies are also outlined.

7.2 Timeline

A Gantt chart is attached in Appendix B outlining the timeline of this project. Deliverables and milestones are included in this timeline.

7.3 Resources Required

The resources that will be used for this project are listed below:

7.3.1 Software Resources

- Stanford Phrasal SMT toolkit
- Fairseq NMT toolkit
- KenLM Language Modelling Toolkit
- OpusTool (Used to format JW300 parallel corpora)
- Google Collab

7.3.2 Hardware Resources

- Computer with standard computing facilities for UCT honours level work
- CHPC cluster

7.3.3 Data Resources

- JW300 English-isiZulu and English-isiXhosa parallel corpora
- Autshmato parallel and monolingual corpora
- MeMaT parallel and monolingual corpora
- Opus Corpus

7.4 Deliverables

The following table lists the final deliverables for this project.

Date Due	Deliverable
4th June	Literature Review
21th June	Project Proposal
23th - 24th June	Project Proposal Presentation
17th Sep	Project Paper Final Submission
20th Sep	Project Code Final Submission
4th - 8th Oct	Final Project Demonstration
11th Oct	Project Poster
18th Oct	Project Web Page

7.5 Milestones

The following table lists the milestones for this project. These milestones are included in the Gantt chart attached in the appendix:

Date Due	Milestone
4th June	Submission of Review
21th June	Submission of Project Proposal
9th July	Submission of Project Proposal Presentation
19th July	Completion of training data sourcing and pre-processing
10th - 13th Aug	Initial Software Feasibility Demonstration
30 Aug	Completion of model training and testing
6 Sep	Submission of Draft Project Report
17th Sep	Submission of Final Project Paper
20th Sep	Submission of Project Code Final
4th - 8th Oct	Final Project Demonstration
11th Oct	Project Poster
18th Oct	Project Web Page

7.6 Work Allocation

The team will collaborate to source and pre-process the necessary training data and design the experiment constraints and specifications. Team collaboration will also be used for shared deliverables such as the project website and poster.
The core experimentation will be done independently, with each team member implementing one of the machine translation approaches ie. SMT or NMT to be used for testing. Each approach will test both data augmentation techniques and the baseline technique. Additionally, each team member will produce a final paper independently.

References

[1] South Africa's people | South African Government.

[2] Top Ten Internet Languages in The World - Internet Statistics.

[3] Aharoni, R., Johnson, M., and Firat, O. Massively Multilingual Neural Machine Translation. *arXiv:1903.00089 [cs]* (July 2019). arXiv: 1903.00089.

[4] Alexander, N. *The potential role of translation as social practice for the intellectualisation of African languages*. PRAESA Cape Town, 2010.

[5] Banerjee, T., Kunchukuttan, A., and Bhattacharya, P. Multilingual Indian Language Translation System at WAT 2018: Many-to-one Phrase-based SMT. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation* (Hong Kong, 2018), Association for Computational Linguistics.

[6] Bertoldi, N., and Federico, M. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation - StatMT '09* (Athens, Greece, 2009), Association for Computational Linguistics, p. 182.

[7] Deng, L., and Liu, Y. *Deep learning in natural language processing*. Springer, 2018.

[8] Dong, D., Wu, H., He, W., Yu, D., and Wang, H. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (2015), pp. 1723–1732.

[9] Duh, K., McNamee, P., Post, M., and Thompson, B. Benchmarking Neural and Statistical Machine Translation on Low-Resource African Languages. In *Proceedings of the 12th Language Resources and Evaluation Conference* (Marseille, France, May 2020), European Language Resources Association, pp. 2667–2675.

[10] Escribe, M. Human evaluation of neural machine translation: The case of deep learning. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)* (2019), pp. 36–46.

[11] Freitag, M., and Firat, O. Complete multilingual neural machine translation. *arXiv preprint arXiv:2010.10239* (2020).

[12] Green, S., Cer, D., and Manning, C. D. Phrasal: A toolkit for new directions in statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation* (2014), pp. 114–121.

[13] Gu, J., Hassan, H., Devlin, J., and Li, V. O. K. Universal Neural Machine Translation for Extremely Low Resource Languages. *arXiv:1802.05368 [cs]* (Apr. 2018). arXiv: 1802.05368.

[14] Ha, T.-L., Niehues, J., and Waibel, A. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798* (2016).

[15] Imankulova, A., Sato, T., and Komachi, M. Filtered Pseudo-parallel Corpus Improves Low-resource Neural Machine Translation. *ACM Transactions on Asian and Low-Resource Language Information Processing 19*, 2 (Mar. 2020), 1–16.

[16] Karakanta, A., Dehdari, J., and van Genabith, J. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation 32*, 1-2 (June 2018), 167–189.

[17] Khayrallah, H., and Koehn, P. On the impact of various types of noise on neural machine translation. *arXiv preprint arXiv:1805.12282* (2018).

[18] Koehn, P. *Statistical machine translation*. Cambridge University Press, Cambridge ; New York, 2010. OCLC: ocn316824008.

[19] Koehn, P., and Knowles, R. Six Challenges for Neural Machine Translation. *arXiv:1706.03872 [cs]* (June 2017). arXiv: 1706.03872.

[20] Lambert, P., Schwenk, H., Servan, C., and Abdul-Rauf, S. Investigations on Translation Model Adaptation Using Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation* (Edinburgh, Scotland, July 2011), Association for Computational Linguistics, pp. 284–293.

[21] Martinus, L., and Abbott, J. Z. Benchmarking Neural Machine Translation for Southern African Languages. *arXiv:1906.10511 [cs, stat]* (June 2019). arXiv: 1906.10511.

[22] McKellar, C., Griesel, M., and Wilken, I. Syntactic Reordering as Preprocessing Step in Statistical Machine Translation of English to Setswana: a linguistically-motivated approach.

[23] Muischnek, K., and Müürisep, K. Impact of corpora quality on neural machine translation. In *Human Language Technologies–The Baltic Perspective: Proceedings of the Eighth International Conference Baltic HLT 2018* (2018), vol. 307, IOS Press, p. 126.

[24] Nyoni, E., and Bassett, B. A. Low-Resource Neural Machine Translation for Southern African Languages. *arXiv:2104.00366 [cs]* (Apr. 2021). arXiv: 2104.00366.

[25] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038* (2019).

[26] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (2002), pp. 311–318.

[27] Ramesh, A., Balavadhani parthasa, V., Haque, R., and Way, A. Investigating Low-resource Machine Translation for English-to-Tamil. In *Proceedings of the*

3rd Workshop on Technologies for MT of Low Resource Languages (Suzhou, China, Dec. 2020), Association for Computational Linguistics, pp. 118–125.

[28] Rubino, R., Marie, B., Dabre, R., Fujita, A., Utiyama, M., and Sumita, E. Extremely low-resource neural machine translation for Asian languages. *Machine Translation 34*, 4 (Dec. 2020), 347–382.

[29] Sennrich, R., Haddow, B., and Birch, A. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709* (2015).

[30] Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* (2015).

[31] Sennrich, R., Haddow, B., and Birch, A. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin, Germany, Aug. 2016), Association for Computational Linguistics, pp. 86–96.

[32] Wu, H., Wang, H., and Zong, C. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08* (Manchester, United Kingdom, 2008), vol. 1, Association for Computational Linguistics, pp. 993–1000.

Appendix

A Project Risks

Risk	Probability	Impact	Consequence	Mitigation	Management	Monitoring
A team member contracts covid-19/is unable to continue with project momentarily or permanenantly	7	5	Increased workload on other team members for shared deliverables.	Avoid engaging in social behaviour leading to exposure to covid-19. Ensure work on shared deliverables begins as early as possible.	Engage with supervisor as soon as illness occurs. Redistribute work on shared deliverables to other team members, according to strengths.	Meet regularly and communicate honestly with regards to health status. Flag and suspected issues early.
A team member's equipment malfunctions during training/testing stage	5	8	Delayed progress of project due to an inability to work.	Backup all project work on a cloud platform allowing access from different devices.	Make use of facilities provided by Computer Science department. Where speciality equipment is required engage with teammates about using their equipment when possible.	Address suspected issues with equipment early and ensure software is up to date.
Chosen toolkits prove insufficient for the requirements of the project	8	8	Project capabilities fail to fulfil project goals/answer research questions.	Research chosen toolkits thoroughly to ensure they are sufficient for use. Make use of supervisor advice and wisdom in assessing appropriateness of toolkit for project use. Begin project work early enough to allow for a software pivot without delaying project timelines.	Where toolkit capabilities are insufficient look for 3rd party libraries/plugins to achieve desired results. Identify alternative toolkits for use in case no such plugins exist.	Begin project work early so as to identify strengths and weaknesses of toolkits quickly.
Long training times	6	8	Unexpected expansion in time allocated for training data leading to inability to meet final project deadlines.	Allocate a substantial amount of time for training and testing purposes so as to account for possible delays. Where necessary make use of additional computation resources to minimise possibility of delay.	Reduce training data so as to quicken training times. Source additional computation resources allowing for increased training times.	Track training time with relation to scheduled training time so as to identify potential bloating in timeline early.
Failure to meet project requirements on time.	6	9	Facing marking penalties.	Stay on schedule as outlined in project timeline. Prioritise core project tasks, in order to avoid scope creep.	Upon discussion with supervisor, reduce project scope so as to allow for on time completion of the project.	Regularly check project progress against project timelines.
Lack of adequate skills to complete the project	4	9	Inability to complete project on time and at a high quality.	Engage with the necessary research and learning materials to develop proficiency in required skills.	Where a skills inadequacy is identified, make use of an online course to improve skills. Additionally, leverage supervisor for his knowledge on project subject matter and guidance towards learning materials.	Begin work on software requirements of the project as soon as possible. Monitor comfort with necessary software tools honestly and continuously.

Table 3: Project Risk Matrix

B Project Timeline

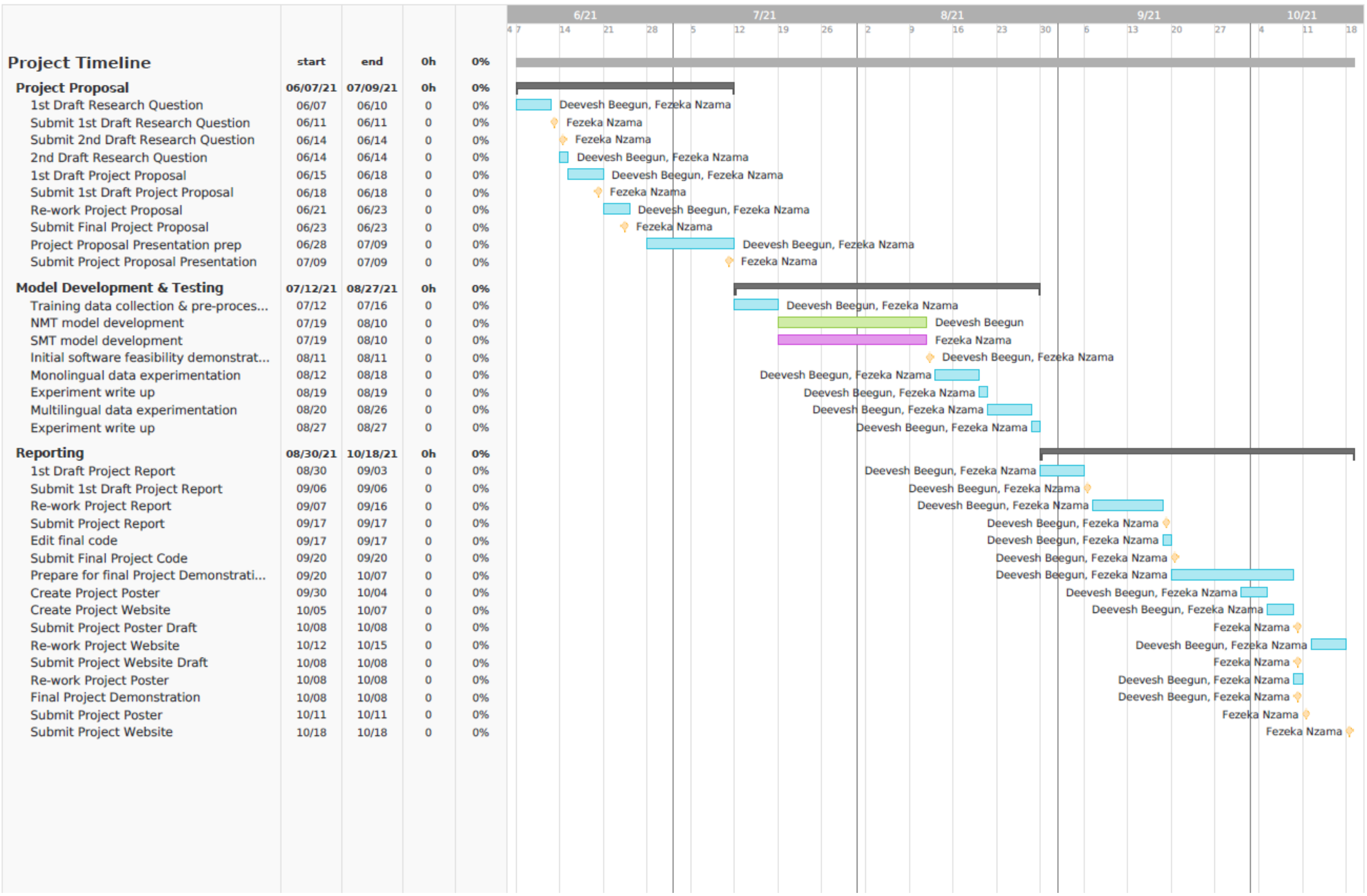

Figure 1: Gantt Chart Showing Detailed Project Timeline