Towards Improving Neural Machine Translation Systems for Lower-Resourced Languages:

Optimising Preprocessing and Data Augmentation Techniques for English to Irish Translation

Joshua Wolfe Brook

Amsterdam University College - Science Major

joshua.brook@student.uva.nl

7 June 2023

Word Count: 8,864

Supervisor:	Dr. Jelke Bloem	UvA	j.bloem@uva.nl
Reader:	Dr. Wilker Aziz	UvA	w.ferreiraaziz@uva.nl
Tutor:	Dr. Maurits de Klepper	AUC	m.c.deklepper@auc.nl

Abstract

Neural Machine Translation (NMT) systems have recently achieved significant advancements in the quality and availability of automatically translated text. However, the lack of readily-available high-quality parallel training data poses a challenge to the robustness and generalisation capabilities of NMT models trained to translate Lower-Resourced Languages (LRLs). This research aims to improve upon these automatic language translation systems, specifically for LRLs, through optimising the data preprocessing pipeline and artificially increasing the size of the training dataset through algorithmic data augmentation techniques. The English - Irish language pair is chosen and used as a case study to determine the effectiveness of the proposed approach, due to the author's familiarity with it and the fact that it should pose a suitable challenge for LRL-based NMT. The proposed system combines various NLP techniques and models, creating an augmentation process which involves targeted word replacement based on POS tags, contextual similarity, and semantic relationships. The newly-generated data is concatenated with the existing data, and used to train multiple versions of a Transformer-based NMT model, in order to compare how training on different datasets can affect translation quality. Both quantitative and qualitative results demonstrate that the preprocessing pipeline and the data augmentation approach improve NMT performance, as evidenced by increased automatic evaluation scores and perceived translation accuracy. These findings highlight the potential of data-based approaches as valuable techniques to enhance NMT systems and address limitations caused by a lack of training data.

Keywords: Low-resource language, neural machine translation, data augmentation, preprocessing, transformer architecture, Irish

Table of Contents

Abbreviations

NLP	Natural Language Processing
MT	Machine Translation
NMT	Neural Machine Translation
LRL	Low-Resource Language
ML	Machine Learning
NN	Neural Network
RNN	Recurrent Neural Network
BERT	Bidirectional Encoder Representaions from Transformers
BLEU	BiLingual Evaluation Understudy
BEER	BEtter Evaluation as Ranking
SGD	Stochastic Gradient Descent
RMSProp	Root Mean Square Propagation
DCEP	Digital Corpus of the European Parliament
DGT-TM	Directorate-General for Translation multilingual Translation Memory
POS	Part-of-Speech
WMT	Workshop on Statistical Machine Translation

Introduction

The overwhelming majority of current natural language processing (NLP) research focuses on only 20 of the world's 7000+ languages, leaving the rest underrepresented, understudied, and underappreciated (Magueresse et al., 2020). To address this divide, the goal of this research is to improve the standards of automatic language translation systems, with a specific focus on lower-resourced languages (LRLs) where major NLP techniques frequently fall short, due to a lack of linguistic resources (Ranathunga et al., 2021). LRLs have a lack of both available data, specifically parallel corpora for training translation models, and language-specific linguistic tools, like tokenisers and part-of-speech (POS) taggers. In the last few years, the field of machine translation (MT) has seen a major shift away from conventional phrase-based translation systems towards statistical algorithmic systems based on neural network architecture (Bahdanau et al., 2015). Typical Neural Machine Translation (NMT) models employ an encoder to convert input sequences into hidden representations and a decoder to unravel those representations and produce sentences in the target language (Bahdanau et al., 2015). Google's Translator, which is still regarded as one of the leading translation systems in the world, has recently transitioned to a neural network-based system, which has proven to have both advantages and disadvantages (Wu et al., 2016). As neural networks are very good at finding hidden patterns in large amounts of data, they can be leveraged effectively in MT tasks when large parallel corpora are readily available for training, however, when attempting to create a translation system involving LRLs, procuring enough useable data can be challenging, meaning that traditional phrase and rule-based approaches can often produce better results in these cases (Ranathunga et al., 2021). [1]

[1] Sections of this paragraph have been taken directly from the author's previous work, completed for the Academic Research Writing course taught at Amsterdam University College (Brook, 2022).

Due to this limited availability of parallel training data, NMT models often face challenges in handling diverse inputs and maintaining robustness, reducing translation quality (Ranathunga et al., 2021). To address this issue of data scarcity that plagues modern NMT methods when training models for LRLs, this research attempts to quantify the impact that two data-related techniques, preprocessing and data augmentation, can have on translation quality of NMT systems, while avoiding any novel approaches to model design. I propose that the data collection, preprocessing, cleaning, and augmentation pipeline can provide just as big an impact on translation quality as changes to model architecture and training methods.

Preprocessing steps play a significant role in improving the performance of NMT systems for LRLs, through text standardisation, noise reduction, and linguistic variation homogenisation, which all allows the model to better understand the fundamentals of each language and translate between them (Koehn, 2010). As Irish is an LRL, much of the available data is noisy, poorly aligned, or poorly translated, meaning extra steps will have to be taken to adequately prepare these datasets (Lankford et al., 2022). Additionally, morphological analysis and disambiguation techniques specific to the target language can assist in improving translation accuracy by resolving ambiguities and capturing correct linguistic structures (Ash et al., 2018). By carefully designing and applying preprocessing methods tailored to the characteristics of both the specific LRL and the specific data, the input data can be prepared in a way that maximises the performance of the NMT system.

By artificially augmenting the training data through techniques such as back-translation, where a parallel corpus is generated by translating monolingual data, the model can be exposed to a wider range of words, structures, and linguistic variations (Fadaee et al., 2017). This data augmentation helps improve the model's robustness, generalisation, and ability to handle diverse sentence structures and

vocabulary (Fadaee et al., 2017). As part of the augmentation, words can be substituted, inserted, and deleted, in order to introduce controlled noise into the data, making the model more resilient to errors and noise in novel sentences (Li et al., 2022).

This research also addresses the topic of automatic translation evaluation, comparing the *de facto* standard method, the BLEU score, proposed by Papineni, Roukos, Ward, and Zhu (2002), against BEER, a more recent alternate evaluation metric, proposed by Stanoyević and Sima'an (2014) to address some of the perceived issues with BLEU.

Overall, the objective of this research is to assess how different data acquisition, preprocessing, and data augmentation methods impact the quality of a Transformer-based NMT system for the specific language pair of English to Irish, where Irish has been chosen as the target language due to both the author's familiarity with it and the fact that it should pose a suitable challenge for MT, as it has major grammatical and morphological differences from English, while belonging to the same broader language family and using the same writing system.

This research is significant, as improving translation quality of low-resource languages can have a positive impact on language preservation and cultural exchange (Ranathunga et al., 2021). The paper's findings will contribute to the development of more accurate and effective automatic language translation systems, benefiting research in the field of NMT and of the Irish language, by addressing methods for improving LRL-NMT systems which are conceptually understandable, computationally feasible, and environmentally conscious (Strubell et al., 2019, Ash et al., 2018, Li et al., 2022).

Irish as an LRL

Irish (*Gaeilge*), sometimes called Irish Gaelic, is a member of the Goidelic branch of Celtic languages, sharing a recent common ancestor with Scottish Gaelic (*Gàidhlig*) in Scotland and Manx (*Gaelg*) in the Isle of Man. It is one of the oldest languages still in use in Europe, with roots dating back more than 2,000 years and a writing system (Ogham) which predates the Latin orthography (Mac Giolla Chríost, 2004). Linguistically, Irish is most well-known for its unique initial consonant mutations (lenition, eclipsis, and aspiration), which change the sound of certain words depending on their grammatical context (Conroy, 2008).

Irish is recognised as the national and first official language of Ireland, but its usage has greatly declined, due in great part to the years of British colonial rule over the island, during which use of the language was actively prohibited (Mac Giolla Chríost, 2004). As of a 2022 census, 71,968 speakers use Irish on a daily basis outside of the education system, the majority of whom are concentrated in small regions, called *Gaeltachtaí*, mostly situated along the western coast of the island (CSO, 2023). Efforts have been made to revitalise use of the language, mostly through the education system and language planning, where it is taught in schools and used in official government communications, but there is still much work to be done to revive its daily use in the majority of the country (Mac Giolla Chríost, 2004).

Advancing NLP standards and optimising NMT systems for Irish is crucial to preserve the language in the digital age (Lynn, 2022). Encouraging accessibility and computational feasibility of the improvement of NMT models for the language encourages research and interest in Irish, thereby encouraging daily usage (Lynn, 2022). Through improving NLP resources and tools for Irish, we can facilitate easier inter-

action with the language through digital platforms, mobile applications, and online services (Lynn, 2022). This accessibility not only improves freedom of information and services for Irish speakers, but may also encourage more individuals to engage with Irish in their daily lives.

A current lack of interest in the language combined with a relatively smaller number of speakers contributes to the scarcity of data and NLP tools for Irish and causes challenges in developing robust NMT models (Lynn, 2022). It is therefore crucial to attract more scholars and researchers to the field in order to create more linguistic resources, curate high-quality parallel corpora, and develop tailored NMT systems that cater to the specific characteristics of the language. Emphasizing the unique linguistic features of Irish could spark interest among linguists, NLP experts, and the average person. By promoting Irish as an intriguing and challenging language to study and develop NMT systems for, we can encourage more academic and industry involvement, ultimately leading to enhanced resources and technologies (Lynn, 2022).

Furthermore, the relationship between Irish and other LRLs, especially related ones such as Scottish Gaelic and Welsh, deserves attention. These languages share historical, cultural, and linguistic connections within the broader Celtic language family, meaning insights and methodologies can be shared and adapted for the advancement of NMT systems for all of these languages (Joshi et al., 2020). Collaboration and knowledge exchange among researchers working on different low-resource languages can result in synergistic efforts that benefit the entire language community (Ranathunga et al., 2021).

Research Context

NMT is a novel method for MT that uses artificial neural network (NN) models to convert input sentences from one language to another (Bahdanau et al., 2015). A typical NMT model comprises an encoder, which converts input sequences into hidden representations, and a decoder, which unravels those representations and generates sentences in the target language (Vaswani et al., 2017). In recent years, NMT has risen to prominence as the preferred method of automatic translation, outperforming traditional statistical and phrase-based approaches (Bahdanau et al., 2015). However, despite its success, NMT translations for LRLs continue to lag behind their high-resource counterparts, largely due to the limited availability of extensive parallel corpora for these underrepresented languages (Ranathunga et al., 2021).

To address this issue, Ranathunga, Lee, Skenduli, Shekhar, Alam, and Kaur (2021) conducted a comprehensive survey of advancements in LRL-based NMT, which offers a valuable resource for identifying techniques to optimise NMT models for any language pair. This article offers a comprehensive overview of recent progress in LRL-NMT research, including an in-depth examination of various advancements and a quantitative analysis to determine the most prevalent solutions. Drawing on insights from prior studies, this survey paper provides practical guidelines for selecting appropriate NMT techniques in low-resource language data scenarios. Moreover, it presents a comprehensive perspective on the research landscape of LRL-NMT and provides recommendations to further enhance research endeavors in this field.

NMT for Irish

Dowling, Lynn, Poncelas, & Way's (2018) paper presents a preliminary comparison between statistical machine translation (SMT) and NMT for English to Irish translation in the public administration domain, and represents the first attempt to apply NMT methods to English-Irish MT. The authors address concerns regarding whether LRLs like Irish can benefit from NMT to the same extent as higher-resourced languages, especially when the NMT system is no longer domain specific (Dowling et al., 2018). The previous success of a domain-specific SMT system in an official Irish government department can generally be attributed to the availability of high-quality parallel data in the domain of public administration, due to extensive historical backlogs and manually-translated documents, but NMT systems face much greater challenges when dealing with LRLs, especially in regards to low-frequency words, long sentences, and out-of-domain subjects (Dowling et al., 2018). By exploring the potential of NMT in the domain of public administration for a low-resource language like Irish, the paper aims to contribute to the advancement of MT for Irish and ensure that it benefits from recent developments in the field as much as possible.

Meanwhile, Lankford, Afli, & Way's (2022) study delves into the impact of hyperparameter tuning on Transformer-based NMT for the English-Irish pair, comparing the model with a Recurrent Neural Network (RNN)-based system and using the BLEU score as a metric for evaluating the accuracy of the translation. They find that the Transfomer model consistently performs better than the model based on an RNN, no matter the size of the training dataset (Lankford et al., 2022). This paper provides some good baselines for translation quality of the Irish-English language pair, which can be used to compare against my own.

WMT18-based Filtering

One of the tasks at the 2018 annual Workshop on Statistical Machine Translation (WMT), organised by StatMT (2018), was a corpus filtering exercise aimed at addressing the challenge of filtering noisy and low-quality parallel corpora for machine translation. The data used is a one billion words parallel corpus for the English-German language pair, taken from a version of Paracrawl, a publically available web-crawled and machine translated parallel corpus project, so the methodology should transfer well to my own research, which uses the same system.

Ash, Francis, & William's (2018) entry to this task presents a novel system for filtering and scoring sentences in a parallel corpus, consisting of two passes: an initial pass which applies hard rules to eliminate low-quality data and a second pass that uses heuristics to assign scores to the remaining sentence pairs. The hard rules filter the sentences based on matching line length, non-translation, language identification, character filtering, and digit matching (Ash et al., 2018). These rules aim to remove sentences with widely varying lengths, untranslated or partially translated sentences, sentences in incorrect languages, unwanted characters, and differing digits (Ash et al., 2018). The scoring heuristics used in the second pass include sentence length, perplexity, and diversity. The sentence length heuristic encourages longer sentences, noting that shorter sentences tended to be of worse quality, while the perplexity heuristic measures the overall perplexity statistics of a "clean" corpus and assigns scores based on similarity to those statistics (Ash et al., 2018). The diversity heuristic measures the similarity between sentences using a rolling buffer and assigns scores based on the sentence's similarity to its neighboring sentences, penalising near-duplicate sentences, as they don't add new information into the dataset (Ash et al., 2018).

Another entry to the WMT18 corpus filtering task was submitted by R. Wang, Marie, Utiyama, & Sumita (R. Wang et al., 2018), who present a similar approach to Ash et al. (2018), performing aggressive filtering to remove obviously noisy sentence pairs, then computing informative features for the remaining sentence pairs - including NMT transformer model scores, lexical translation probabilities, and bilingual word embeddings. R. Wang et al.'s (2018) primary aggressive filtering method removes any sentence pairs where more than 25% of the tokens are numbers or punctuation marks, or which are very long or very short. A logistic regression classifier was trained using positive and negative examples to compute scores for each sentence pair, which were used to rank the sentence pairs and select the top-ranked sentences for building a robust NMT system (2018). The authors evaluate the performance of their NMT systems trained on the filtered data, showing that both the aggressive filtering and the classifier-based filtering improved the performance of the model by more than 20 BLEU (R. Wang et al., 2018).

These two systems (Ash et al., 2018, R. Wang et al., 2018) aim to provide a general approach to noisy corpus filtering that can be applied to other language pairs and datasets, rather than being specifically tuned for the task at hand. They produce competitive results without excessive fine-tuning and offer a fast initial filtering step that already provides significant gains (Ash et al., 2018, R. Wang et al., 2018).

Data Augmentation

Data augmentation is a technique commonly used in ML to increase the size and diversity of a small training dataset by artificially creating new data instances (Quteineh et al., 2020). Typically, it involves applying various transformations or modifications to the existing data while preserving the original label or target information. For example, in data augmentation for computer vision, images can often be cropped, flipped, or rotated while preserving their original content, and therefore, their original label (Fadaee et al., 2017). However, as Fadaee, Bisazza, and Monz (2017) demonstrate in their paper on Data Augmentation for LRL-NMT, the augmentation of parallel corpus data needed for NMT is more difficult, as in most cases, both source and target sentence must be altered.

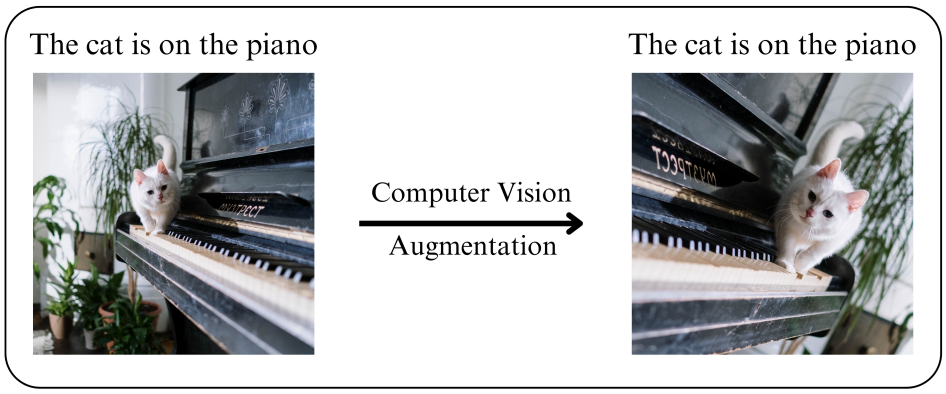

Figure 1: Label-preserving flip, rotate, and crop data augmentation for computer vision (Pexels, 2021)

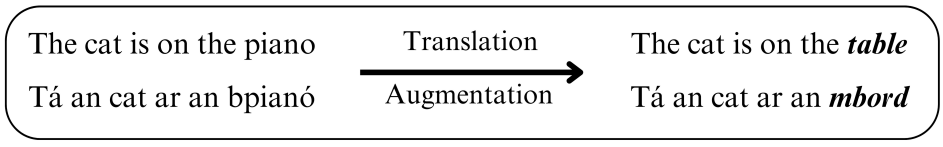

Figure 2: Both source and target sentence must be altered in data augmentation for parallel corpora[2]

[2]Above figures adapted from Fadaee et al. (2017)

A paper by X. Wang, Pham, Dai & Neubig (2018) presents a novel method for NMT data augmentation, formulating the design as an optimisation problem and proposing a simple and effective augmentation strategy they call SwitchOut. This method focuses on randomly replacing words (and their translation) in both the source and target sentences with other random words (and translations) from their respective vocabularies (X. Wang et al., 2018). The authors show that SwitchOut consistently improves translation performance compared to other augmentation techniques like word dropout (X. Wang et al., 2018). They introduce two key properties that the augmentation policy should possess: diversity, meaning that the augmented data points should be more diverse than the observed training data, and smoothness, meaning that similar sentences should have similar probabilities (X. Wang et al., 2018).

A generic framework for data augmentation is described, encompassing existing methods such as word dropout and Reward Augmented Maximum Likelihood (RAML), where the similarity between augmented and observed pairs is measured using domain-specific similarity functions for source and target sentences (X. Wang et al., 2018). Practically, this procedure involves randomly selecting words for replacement in the source and target sentences, based on a negative Hamming distance distribution (X. Wang et al., 2018). The SwitchOut method is shown to be simple, efficient, and competitive with existing augmentation techniques, validated through experiments on three translation datasets, where results demonstrate consistent improvements of 0.5 BLEU in translation quality compared to strong baselines, including word dropout (X. Wang et al., 2018).

A survey paper written by Li, Hou, and Che (2022) expands on the methodology proposed by X. Wang et al. (2018), exploring a paraphrasing-based method which aims to generate augmented data with limited semantic difference from the original data.

The word-level paraphrasing methodology aims to identify specific keywords in a sentence, and replace them dynamically through the use of thesauruses, semantic word embeddings, and large language models (LLMs) (Li et al., 2022). Thesauruses, e.g. WordNet, are lexical resources that provide manually-curated synonyms, antonyms, and hypernyms for given words, while semantic word embeddings, e.g. Word2Vec, are representation of words or phrases in a numerical vector space which capture the semantic and syntactic relationships between words based on their contextual usage in a given corpus of text (2013).

In essence, words with similar meanings or related contexts should have similar vector representations, and as such, have small cosine similarities between each other in the given vector space (2013). Geometric properties of vectors enable mathematical operations on word embeddings. For example, vector addition and subtraction can capture relationships like `king - man + woman = queen`, showcasing analogy reasoning capabilities, which is useful for targeted-word-replacement-based data augmentation (2013).

Pretrained LLMs, e.g. BERT (Bidirectional Encoder Representations from Transformers) and ChatGPT (OpenAI, 2023), have revolutionised the entire field of NLP, and can be used to predict masked words in text based on context as well as generate human-like text with a high degree of fluency (Li et al., 2022). BERT is based on the Transformer architecture, which consists of multiple layers of self-attention mechanisms and feed-forward neural networks (Vaswani et al., 2017). During training, the model learns to predict missing words in sentences and determine whether two sentences are consecutive paired (Vaswani et al., 2017). This makes BERT especially good at capturing a full bidirectional context of words and sentences, resulting in a deeper understanding of language semantics (Vaswani et al., 2017).

As these paraphrasing techniques offer different levels of flexibility and control over the generated paraphrases, it's worth noting that these methodologies can be combined or customised to suit specific needs. For instance, a combination of word embeddings and LLMs can be used, where word embeddings are employed to select candidate replacement words, and language models are used to fine-tune generated paraphrases (Mikolov et al., 2013). This idea will play a key role in the development of my own data augmentation system, described later in this paper.

Fadaee et al.'s (2017) paper proposes a data augmentation approach called Translation Data Augmentation (TDA) to improve LRL-NMT. Inspired by data augmentation techniques in computer vision, the authors apply a similar concept to NMT by generating new sentence pairs that contain rare words in synthetic contexts (Fadaee et al., 2017). The goal is to provide novel contexts for rare words, as they are challenging to model accurately and are more prevalent in low-resource settings, when there is data scarcity (Fadaee et al., 2017). The proposed method involves selecting rare words, replacing them in the original sentences using a pretrained language model, and selecting corresponding translations based on word alignments (Fadaee et al., 2017). Experimental results on simulated low-resource settings demonstrate that TDA improves translation quality by up to 2.9 BLEU points over a defined baseline and up to 3.2 BLEU points over a back-translation-based system (Fadaee et al., 2017).

Quteineh, Samothrakis, & Sutcliffe (2020) suggest that using guided outputs of a language generation model, such as GPT-2, to generate additional data that complements the existing dataset could improve the performance of text classifiers through an active learning (see Settles, 2011) process. They transform the data generation task into an optimisation problem, using Monte Carlo Tree Search (MCTS) as the optimisation strategy and incorporating entropy as one of the optimisation criteria (Quteineh et al., 2020). The active learning process starts with a small labeled

dataset, which the proposed approach leverages to generate synthetic examples with MCTS, guiding the language generation model to output informative examples, which are manually labeled and added to the training set (Quteineh et al., 2020). The authors compare their approach with a Non-Guided Data Generation (NGDG) process, where the data generation is not optimised for a reward function, showing that the proposed approach with MCTS achieves increased performance of up to 26% compared to NGDG on two different datasets (Quteineh et al., 2020). Through generating synthetic data and guiding the process using MCTS, the approach shows promising results in improving text classifier performance (Quteineh et al., 2020). As will be discussed further in this paper, the capability of LLMs to both augment and generate parallel corpus data are impressive, and should be fully utilised to improve the field of NMT, especially for LRLs

Overall, these papers all highlight the potential of data augmentation to address the challenges faced by LRL pairs in NMT, and provide a plethora of potential methods for creating my own data augmentation system which suits the English-Irish pair.

Computational & Environmental Considerations

As discussed in Cheng, Wang, Zhou, Zhang's (2020) survey, training large-scale NMT models typically requires substantial computational resources, including the use of powerful GPUs or TPUs and access to immense amounts of memory. The cost associated with accessing or acquiring and maintaining such infrastructure is prohibitively expensive, particularly for individuals with limited resources (Cheng et al., 2020). Even with large resources, training full-scale NMT models can take days, weeks, or even months to complete a single training run, especially as datasets get bigger and model architectures more complex (Strubell et al., 2019).

Training large NMT models also consumes significant amounts of energy, contributing to the ever-growing carbon footprint of Artificial Intelligence (AI) research and development (Strubell et al., 2019). In regards to Transformers in particular, Stubell, Ganesh, & McCallum (2019) estimate that to train a base Transformer model (65M parameters) on one Nvidia P100 GPU for 12 hours produces roughly one kilo of CO_2 and would cost between \$5 and \$17 of cloud compute time on Amazon Web Services, currently the most popular cloud computing service. The environmental impact of energy-intensive computations is a growing concern, and reducing energy consumption is essential for sustainable AI development (Strubell et al., 2019).

In summary, the computational requirements for training large-scale NMT models pose significant challenges, especially for individuals, researchers, and organisations with limited computational or financial resources, while the ever-growing energy consumption associated with training NMT models contributes to the already very large carbon footprint of AI, meaning innovative solutions to minimise computational requirements will be essential to make NMT research and development more accessible and environmentally friendly (Strubell et al., 2019, Cheng et al., 2020).

Methodology

Data Collection

As previously mentioned, LRLs are defined by having a lack of available data, in particular for NMT, parallel corpora - sentences in both languages which have been translated by a human, which is still seen as the gold standard for translation tasks. As such, the first challenge is to collect parallel training data for the English-Irish language pair. In this section, the five parallel corpora used to train versions of the NMT model are presented and compared.

Tatoeba (2006) provides a large volume of accurate translated sentences in various languages, but unfortunately only has 2000 sentences for the Irish–English pair, far less than the minimum recommended amount of 10000 suggested by Joshi, Santy, Budhiraja, Bali, and Choudhury (2020) in their comprehensive overview of the state of LRLs in NLP. Joshi et al.'s (2020) research lists Irish as a class 2 language, meaning there exists both a small amount of accurately labelled data and a support community actively trying to keep the language relevant and accessible on a global scale. This lack of accessible data is the main issue facing translation involving LRLs with most of the proposed models, and one which may be combated by use of a pretrained model (Li et al., 2022), or of course, data augmentation.

The website *nlp.irish* (2020) provides a list of useful resources for NLP involving Irish, including links to several parallel corpora. Two such datasets are the Digital Corpus of the European Parliament (DCEP) (2014), which includes a wide range of bilingual documents (mostly press releases, session records, and legislative documents) sourced from the official website of the European Parliament, covering the majority of their

publications between 2001 and 2012, and the European Commission's Directorate-General for Translation's *Multilingual Translation Memory* (DGT-TM) (2012), which is comprised of all of the legislative documents of the European Union, in all 24 official languages of the EU, which Irish has been a part of since 2007. Both of these datasets are large and relatively clean, with 40,000 well-translated parallel sentences in each one. However, due to the topic nature of both datasets, the text contained within is full of domain-specific dense legalese, technical language, and some extremely long sentences. As such, training an NMT model on this data may not allow it to generalise well to translation involving more mundane topics (Dowling et al., 2018).

Paracrawl (2020) is a large-scale web crawling and machine translation project that aims to create parallel corpora for multiple language pairs. It utilises web crawling techniques to gather data from the internet and then passes the collected data through a machine translation model to generate a parallel corpus for the collected texts. Using this resource consisting of NMT-translated data to train another NMT model is, of course, not ideal, but this is one of the main issues with LRL-NMT, and one which can be somewhat circumvented through appropriate cleaning and preprocessing. The project focuses on low-resource languages and language pairs for which there is limited availability of parallel corpora. The most recent version (V9) of Paracrawl supports an improved pipeline with better PDF processing, language identification, and neural-network-based cleaning than previous versions (Bañón et al., 2020). Version 9 contains 3,245,618 parallelised sentences for the English - Irish language pair, which, although cleaner than previous versions of the data, will still need a non-trivial amount of fine-tuning.

As mentioned previously, Quteineh et al.'s (2020) paper demonstrates that using pretrained models for data augmentation is highly effective, and manual testing proved that LLMs are not only effective at augmenting prexisting data, but also

for generating parallel corpora from scratch. As such, a corpus of 1000 parallel sentences has been generated using OpenAI's LLM, ChatGPT (2023). The system was prompted to return parallel translations of sentences which cover various topics and use different verbs to provide a diverse range of examples. The system at first had some issues - changing the formatting every few lines, and only generating very simple sentences all only using the verb 'tá' (to be, present tense). However, with some more specific prompting, the dataset was generated efficiently and without much issue. Manual evaluation of a sample of the translations proved that they were of good quality.

Table 1: List of datasets

Dataset	Description	No. Sentences
Tatoeba	Human-curated dataset with a range of topics	2,388
ChatGPT	Bicorpus generated by an LLM	1,000
DCEP	Assorted texts from the European Parliament	46,418
DGT-TM	Legislature of the European Union	90,819
Paracrawl	Web-crawled & Machine Translated Bicorpus	3,092,955

Using modern LLMs trained on large multilingual corpora, such as the GPT models, to generate parallel corpora is an interesting new development, which should be researched further. It is currently relatively inefficient to generate large corpora using the ChatGPT's graphical user interface (GUI) to generate small batches of sentences at a time, but paid access to the API may alleviate this problem and allow for efficient and customisable corpus generation. Finally, it is worth mentioning that it may also be possible to obtain additional data through manual human translation and crowd-sourcing, if necessary, but this research attempts to avoid that, as it can be costly, inefficient, and hard to regularise.

Data Preparation

Preprocessing must be carried out for all of the data, reducing the complexity of the sentences, which may be disorganised, missing sections, or otherwise irrelevant. Overall, preprocessing is an essential step that ensures the training data is of high quality and suitable for training the NMT model (Ranathunga et al., 2021, Ash et al., 2018). The main preprocessing pipeline involves cleaning, tokenisation, and vectorisation.

Cleaning and filtering must be carried out for the Paracrawl data, as it is extremely noisy, containing translation errors, misaligned sentence pairs, and lots of irrelevant data (numbers etc. which don't need to be translated). The cleaning pipeline is constructed in accordance with the processes described by Ash et al. (2018) and R. Wang et al. (2018) in their papers on their respective systems for filtering noisy parallel data, carried out for WMT18. The basic cleaning involves process involves normalisation - converting the text to lowercase and expanding contractions and abbreviations - and removing unwanted characters, such as HTML tags, excessive punctuation, and any non-alphanumeric characters from the text.

Next, as in Ash et al. (2018) and R. Wang et al. (2018), sentences are length-matched, as ensuring that the lengths of the English and Irish sentences are relatively similar is a good preliminary check for an accurate translation. Inappropriate or offensive language is removed from the sentence pairs and language identification is performed on the Irish sentences using the langid Python library. Only sentences where the supposed Irish sentence are actually classified as Irish by langid are retained in the corpus.

Finally, once the noisy Paracrawl data is cleaned, all of the datasets can be tokenised and vectorised. Tokenisation is the process of breaking down the text into smaller

meaningful units, usually whole words or occsaionally individual morphemes. Tokenisation for the English sentences is trivial, using the standard `word_tokenize` method provided by the Python Natural Language ToolKit (NLTK), which is widely-used and accepted to be reasonably accurate for English (Lynn, 2022). Here is another problem that Irish faces as an LRL, a lack of pretrained or language-specific NLP tools, such as tokenisers, lemmatisers, and POS-taggers (Lynn, 2022). These are all useful tools, common in preprocessing pipelines, which tend to be langauge-specific, as all languages have different word order, grammar, and inflectional morphology (Lynn, 2022).

After some experimentation, it was decided that NLTK's `word_tokenize` should perform well enough for the purposes of this research, but is far from optimal. As an example of an issue, since Irish has its initial consonant mutation represented in the orthography, words change depending on their context, (e.g. *bord* (table), *ar an mbord* (on the table)). The standard NLTK tokeniser doesn't know this however, so it thinks *bord* and *mbord* are two different words rather than two versions of the same word. This could impact translation quality as the model may also recognise them as different words.

Using TensorFlow, two `TextVectorization` layers, one for English and one for Irish, are defined. These layers are used to convert the text data into a numerical format that can be fed into the NMT model. These vectors are adapted to the English and Irish data, allowing the layers to learn the vocabulary of the input text data and map each word to a unique integer. A TensorFlow `Dataset` is then created from the two vectors, which is batched, shuffled, and prefetched for improved training performance (Chollet, 2021).

Data Augmentation

Data augmentation is a promising technique which can be used to alleviate issue of a scarcity of parallel training data (Fadaee et al., 2017, Li et al., 2022). By algorithmically generating additional synthetic examples that closely resemble the original data, augmentation helps create a larger, more diverse and more representative training set (Fadaee et al., 2017). In this research, a rule-based approach to data augmentation for NMT is explored, generating novel sentences by replacing specific words while preserving the original sentence structure.

Automatic POS tagging is employed to assign grammatical tags to each word in the English sentences. Both NLTK and SpaCy, two of the biggest python libraries for NLP provide POS-tagging capabilities, but through qualitataive testing, it was found that both had issues with tag accuracy. Hence, Flair was chosen, based on a the results of a quality assessment of major POS-tagging frameworks carried out by the Association for Computational Linguistics (ACL) ("POS Tagging", 2019). For example, when given the sentence "What does a biologist study?" both NLTK and SpaCy POS taggers list "study" incorrectly as a noun ("a study"), while Flair correctly tags it as a verb ("to study"). The choice of targeting specific words based on their POS tags allows for controlled and meaningful replacements that retain grammatical coherence (Li et al., 2022). Through extensive testing and an active-learning-based qualitative analysis, focusing on nouns (NN, NNS, NNP), gerund verbs (VBG), adjectives (JJ), and possessive pronouns (PRP$) gave the most coherent results, only introducing changes to vocabulary, allowing for targeted word replacement which retains structure and grammatical coherence.

Once sentences have been POS-tagged, Word2Vec embeddings are used to identify suitable alternative words for targeted replacement (R. Wang et al., 2018, Li et al.,

2022). These word embeddings are numerical vector representations of words, which capture semantic and grammatical relationships between words based on word co-occurrence patterns found in large text corpora (Mikolov et al., 2013). The Word2Vec model learns to map words into a continuous vector space, enabling the measurement of semantic similarity between words based on their relative positions (Mikolov et al., 2013). This particular model is trained on a large-scale Wikipedia corpus, trained on 2B tweets, and offers a rich semantic representation of words.

In addition to Word2Vec embeddings, a context index is built using the NLTK library. The context index enables the retrieval of similar words based on their context within the training corpus (Mikolov et al., 2013). By leveraging both the Word2Vec embeddings and the NLTK context index, the algorithmic approach ensures that word replacements in augmented sentences are contextually appropriate, thereby maintaining coherence and preserving the intended meaning of the original sentences (R. Wang et al., 2018). By incorporating both Word2Vec embeddings and the context index, the algorithm combines semantic similarity and contextual appropriateness in the word replacement strategy.

During augmentation, the code randomly selects 10 alternative words taken from both the context index and Word2Vec embedding of a given word in the sentence, and shuffles them randomly to introduce variation, ensuring that the augmentation process explores a range of possible replacements. The proposed alternative words are iterated through, and substituted into the sentence, which can once again be POS-tagged, to ensure that the new word matches the target word's POS, as well as being semantically related.

This method of word replacement enables the augmentation process to introduce variations while preserving semantic coherence. By exploring the vector space of

word representations, alternative words with similar semantic meanings but different surface forms can be identified (Mikolov et al., 2013). This variation in word choice enhances the diversity of the augmented data and contributes to the robustness of the NMT model (Li et al., 2022).

After the augmentation process, the modified English sentences undergo final grammar and style checking using the LanguageTool Python library to ensure that grammatical correctness is maintained and stylistic conventions are adhered to. This step is essential to ensure the quality and coherence of the augmented data. Finally, the augmented English sentences can be translated into Irish using Google's Translation API. By leveraging a high-accuracy pretrained NMT model like Google's Translator, new sentence pairs can be created, while ensuring accuracy and reliability of the generated Irish sentences (Wu et al., 2016).

Ideally, data augmentation could be done without relying on an external automatic translation system, and instead substituting translations of words into the Irish sentence from an appropriately accurate bilingual dictionary (Qutineh et al., 2020). However, due to the low-resource nature of Irish, finding an sufficiently extensive, accurate bilingual dictionary which could be downloaded in a readable format proved impossible, and translation through this method proved reasonably effective.

To assess the coherence and translation quality of the newly augmented data, both qualitative and quantitative analyses are performed. Selected examples are manually reviewed to assess the quality and diversity of the augmented translations, considering linguistic accuracy, fluency, and the ability to handle structural variations. Quantitatively, the augmented dataset, comprising of both the original and augmented pairs, is used to train a new NMT model, whose performance can be automatically evaluated through use of the previously mentioned BEER and BLEU scores.

Model Implementation

A baseline Sequence-to-Sequence (Seq2Seq) Transformer-based encoder-decoder model is implemented, based on code provided by the Keras Documentation (Chollet, 2021). This baseline is important to define, as it will provide a standard target of translation quality to test various datsets and data procesing methods with.

This Transformer-based system, based on Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, & Polosukhin's (2017) proposed architecture, extracts representations of words and phrases through encoding into a fixed-length representation which uses self-attention mechanisms to capture dependencies between input tokens. The extracted representation is then passed to a decoder which generates the translated text by predicting the next word in the target language at each step, using both the encoded representation and its own internal state to make predictions (Vaswani et al., 2017)

This Seq2Seq model consists of a `TransformerEncoder`, which processes the input sequence to generate a modified representation, and a `TransformerDecoder`, which takes this representation and the current target sequence (words 0 to N) to predict the next words in the target sequence (up to N+1) (Vaswani et al., 2017). To preserve the order of words, the model incorporates a `PositionalEmbedding` layer. In a standard NN layer, each word in the vocabulary is represented by a vector of fixed dimension. However, this approach does not take into account the position of the word in the sequence. A positional embedding layer addresses this limitation by adding a fixed vector to each word embedding that encodes its position in the sequence (Vaswani et al., 2017). The positional embeddings are learned during training and are added to the word embeddings before feeding them into the model (Chollet, 2021). To ensure that the `TransformerDecoder` only uses information from

tokens up to the current position being predicted, the model applies causal masking (Vaswani et al., 2017). Thus, the `TransformerDecoder` only utilises information from target tokens 0 to N while predicting the next token at position N+1, as it observes the entire sequence at once (Chollet, 2021). Finally, self-attention mechanisms are also used, to allow the model to focus on different parts of the input sequence at each step, depending on their relevance to the current output (Vaswani et al., 2017).

After defining the model, it needs to be trained on the preprocessed data, which involves dividing the data into training and validation sets, and using an optimisation algorithm, in this case, Keras' Adam (Adaptive Moment Estimation), to adjust the parameters of the network in order to minimise the loss function between predicted and desired outputs. The optimisation algorithm aims to to find the optimal set of weights and biases which will allow the neural network to make accurate predictions on unseen data (Ruder, 2016). Adam has become a popular optimisation algorithm, combining both stochastic gradient descent (SGD) and RMSProp (Root Mean Square Propagation) to create a powerful optimiser (Ruder, 2016). As an extension of SGD, at each iteration Adam calculates the adaptive learning rate for each parameter based on the estimated moments (mean and variance) of the gradients (Kingma and Ba, 2014). It then updates the parameter values by moving them in the direction of the negative gradient of the objective function with respect to the parameters (Kingma and Ba, 2014).

Additionally, due to a lack of access to budget, time, or dedicated servers to train models, the use of computational resources will be kept to a minimum, and all model training will be carried out using the 30 hours per week of access to an Nvidia Tesla P100 GPU provided free of charge by Kaggle Notebooks.

Model Evaluation

Automatic evaluation of the quality of automatically generated translations is crucial for the development of high-quality MT systems. The BLEU score, a metric developed by Papineni, Roukos, Ward, & Zhu (2002), has become the *de facto* standard evaluation metric in the field, but has come under increasing scrutiny in recent years. One major issue with BLEU is the observed inconsistency between BLEU scores and manual human evaluations, addressed in papers by Callison-Burch, Osborne, & Koehn (2006) and Tan, Dehdari, & Van Genabith (2015). BLEU primarily focuses on n-gram precision, measuring the overlap between reference and candidate translations, but places no explicit constraints on the order of matching n-grams (Papineni et al., 2002). Therefore, it does not explicitly consider fluency or correctness of translations, only the amount of overlapping words, and as a result, may assign high scores to translations that contain correct n-grams but lack overall coherence or which have grammatical errors (Callison-Burch et al., 2006).

BLEU is also found to be lacking when it comes to understanding intent and meaning in sentences, partly due to its n-gram-focused scoring system, which neglects word order. This can mean two sentences with opposite meanings (e.g. "The dog ate the cat" and "The cat ate the dog") could be scored highly by the metric (Tan et al., 2015). It may also assign lower scores to translations that contain correct information but express it using different vocabulary or phrasing, which hampers the evaluation of translations that deviate from the specific phrasing used in the references (Callison-Burch et al., 2006). To address these issues, many new evaluation metrics have been proposed as alternatives, with some being more suitable than others for certain language pairs and tasks.

The BEER (BEtter Evaluation as Ranking) score, proposed by Stanoyević and Sima'an of the Universiteit van Amsterdam (2014) is one such alternatively-proposed metric which aims to remedy some of the shortcomings of the BLEU score. BEER 2.0 is an evaluation metric that operates at the sentence level and boasts two main advantages: (1) a streamlined method to efficiently tune a vast number of features, maximising correlation with human evaluation, and (2) smoother sentence level scores due to many simple features being combined in a tuneable linear model (Stanojević and Sima'an, 2014).

The BEER evaluator takes two arrays of aligned sentences, one of translations outputted by the NMT model and one of reference translations to compare against, and outputs a corpus-level evaluation score by default. For each model trained on a different dataset, 1000 sample translations are taken and scored by BEER, returning a list of scores which can be compared.

Results

The results obtained from the quantitative analysis demonstrate the potential of the algorithmic data augmentation approach in improving NMT performance. The NMT model trained on the augmented dataset shows a massive improvement in both BEER (55.49) and BLEU (47.36) scores, as well as through manual evaluation of a sample translation accuracy, compared to the model trained solely on the original dataset (BEER=17.41, BLEU=12.59) This indicates that the augmented data enhances the model's ability to handle diverse inputs and improves generalisation.

A considerable portion of this increase in translation quality is of course due to the size of the augmented dataset, which is roughly four times larger than the original. However, through taking a sample of a similar size to the augmented data from the Para-Clean data, we can compare the quality of the two and see that the translation evaluation scores for the Para-Clean sample are much lower (BEER=25.02, BLEU=24.12), meaning that our augmented data is indeed better for training a translation model than the noisy Paracrawl data, even when cleaned.

Table 2: Evaluation Comparison

Dataset	Component Datasets	No. Sentences	BEER	BLEU
Small	Tatoeba, ChatGPT	3,388	17.41	12.59
Aug	Small, Augmented	13,745	55.49	47.36
Para-Sample	Paracrawl V9 Cleaned Sample	10,000	25.02	24.13
Para-Base	Paracrawl v9 Unfiltered	3,092,955	52.34	65.30
Para-Clean	Paracrawl v9 Cleaned	1,476,816	40.14	41.41
Quality	Small, DCEP, DGT-TM	140,625	41.53	43.65
Full	Quality, Para-Clean, Aug	1,627,798	48.39	52.11
Baseline	DGT-Trans-Base[3]	52,000	-	53.40

[3]Baseline Transformer model from Lankford et al. (2022)

However, through qualitative analysis, it can be seen that none of the smaller models (Small, Aug, and Para-Sample) produce "good" translations. with most of the translated results being completely incorrect. The outputs from the Small dataset are grammatically incoherent, getting stuck repeating words and generally not generating anything that could remotely be considered good Irish. The Aug and Para-Sample datasets both perform better, often generating sentences which are grammatical but unrelated to an ideal reference translation

Table 3: Example Translations

Dataset	NMT Model Output	Manual Back-Translation
Manual	Chuaigh mé go dtí an siopa	I went to the shop
Small	Chonaic mé an an ag an an ag...	I saw the the at the the at...
Aug	Tá sé go leor de dhíth ort	You need it a lot
Quality	Chuaigh mé ar an siopa	I went on the shop
Full	Chuaigh mé ar an siopa	I went on the shop
Para-Clean	Chuaigh mé go dtí an siopa	I went to the shop
Para-Base	Chuaigh mé go dtí an siopa	I went to the shop

Both BEER and BLEU scores are relatively consistent for each dataset, with the highest absolute differences appearing at the highest (Para-Base: BEER=52.34, BLEU=65.30, Diff=12.96) and lowest (Small: BEER=17.41, BLEU=12.59 Diff=4.82) scores. However, with this small a sample size, it is challenging to assert which metric performs better. None of the models except for Para-Base exceed the standard of BLEU set by Lankford et al.'s (2022) base Transformer model, as less computational time was spent on each model, which were trained each for a maximum of 12 hours, due to computational constraints. Despite the lack of resources, the results provide evidence that more data is almost always better, and higher volumes of higher quality data are even better. The filtering of the Paracrawl data was ultimately unsuccessful, with the base model providing considerably higher evaluation scores than the cleaned version, which was less than half the size.

Discussion

Due to the improved evaluations scores demonstrated above, this research demonstrates the effectiveness of the proposed data preprocessing pipeline and algorithmic data augmentation techniques in improving the performance of NMT systems for LRLs. Through optimising the preprocessing pipeline, and artificially increasing the size of the training dataset through use of a targeted word replacement algorithm based on POS tags, contextual similarity, and semantic relationships, the NMT model trained on the augmented dataset showed significant improvements in translation quality compared to the model trained solely on the original data. The qualitative and quantitative results demonstrate increased automatic evaluation scores and perceived translation accuracy, highlighting the potential of algorithmic data generation, as well as automatic evaluation methods.

The research builds upon previous studies on corpus filtering and data augmentation for LRL-NMT. The approach to large corpus filtering, inspired by Ash et al. (2018) and Wang et al. (2018), provides a general method for filtering noisy and low-quality parallel corpora, resulting in significant gains in NMT performance. The data augmentation technique, influenced by the approaches of X. Wang et al. (2018), Li et al. (2022), and Qunteineh et al. (2020), leverages a rule-based word replacement system to generate augmented data that closely resembles the original data.

While the proposed approach has shown promising results, there are several limitations to be addressed. The use of an external automatic translation system (Google's Translation API) for generating the Irish translations in the augmented dataset is a potential issue, as it introduces complexity and additional compute time into the augmentation process (due to the volume of API requests). As such, it would be

beneficial to investigate the use of alternative automatic translation systems or use of bilingual dictionaries for generating parallel translations instead of relying on an API, as this could potentially improve the accuracy of the generated Irish sentences while removing the reliance on an external translation system (R. Wang et al., 2018, Li et al., 2022).

Future research should focus on exploring additional augmentation strategies and evaluating their impact on translation on other NMT models, and involving different language pairs and domains. Moreover, fine-tuning the augmentation process and optimising the selection criteria for word replacements may further enhance the quality and diversity of the augmented data (Fadaee et al., 2017). This would provide a more comprehensive understanding of the effectiveness and generalisability of the proposed approach (Li et al., 2022).

As mentioned previously, further research into the capabilities of LLMs to efficiently generate large parallel corpora should be undertaken. As the field of artificial intelligence (AI) continues to improve, effective utilisation of pretrained LLMs for a variety of tasks will become a key area of research to be explored. LLMs as large as, e.g. ChatGPT (OpenAI, 2023), can't possibly be trained by individuals with limited computational resources, but API access to such services could allow the average person to easily generate and augment their own parallel corpora. However, the use of these LLMs is of great concern from an enviromental perspective, with research on BLOOM, a similarly sized model to ChatGPT, estimating that its full training process may have omitted up to 50 tonnes of CO_2 (Luccioni et al., 2022). Developing LLMs more sustainably and providing barrier-free access to their use, while attempting to remain carbon-neutral will become increasingly important in the coming years, and should be a top priority to address for large companies and institutions working in the field (Cheng et al., 2020, Strubell et al., 2019).

Conclusion

In conclusion, this research acknowledges the underrepresentation and underappreciation of the majority of the world's languages in current NLP research, and aims to bridge the gap in the quality of NMT systems for LRLs and emphasise the need for dedicated efforts in the field. The effectiveness of NMT systems depends on the availability of large parallel corpora for training, which poses a major challenge for LRLs where data scarcity is common, historically making traditional phrase-based approaches more favorable. To address this issue, this research demonstrates the potential of data-driven optimisation, including preprocessing and data augmentation, as valuable techniques to enhance NMT systems and address the limitations caused by a lack of training data. By leveraging algorithmic data augmentation and optimising the data preprocessing pipeline, significant improvements in translation quality can be achieved, as evidenced by both quantitative evaluation scores and qualitative analysis.

The results of this study will have a positive impact on the advancement of automatic language translation systems, benefiting both the field of NMT research and the Irish language. The focus on improving LRL-NMT systems aligns with the objective of developing methods that are conceptually understandable, computationally feasible, and environmentally conscious, as highlighted in previous research (Strubell et al., 2019, Ash et al., 2018, Li et al., 2022).

As mentioned previously, to preserve the Irish language in the digital age, it is crucial to advance NLP standards and optimise NMT systems for the language (Lynn, 2022). This involves making NMT models for Irish more accessible and computationally feasible to train and tune, encouraging research and the promotion of daily usage of and genuine interest in the language (Lynn, 2022). By improving

NLP resources and tools, we can facilitate comfortable interaction with Irish, benefiting both fluent speakers and those interested in learning the language (Lynn, 2022.

Focusing specifically on the English-Irish language pair, this research highlights the challenges posed by the morphological and grammatical differences between the two languages. Irish, with its rich historical roots and linguistic uniqueness, serves as a suitable candidate to assess the impact of different data acquisition, preprocessing, and augmentation methods on Transformer-based NMT systems. Finally, through improving upon data-focused LRL-NMT standards, this research contributes to language preservation, cultural exchange, and the development of more accurate and effective automatic language translation systems. It underscores the importance of dedicated research efforts for LRL-NMT and their potential to benefit both NLP research and language enthusiasts alike.

References

Ash, T., Francis, R., & Williams, W. (2018). The Speechmatics Parallel Corpus Filtering System for WMT18. https://doi.org/10.18653/v1/w18-6472

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. https://arxiv.org/pdf/1409.0473

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Rojas, S., Sempere, L. P., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B. J., Waites, W., Wiggins, D., & Zaragoza, J. (2020). ParaCrawl: Web-Scale Acquisition of Parallel Corpora. https://doi.org/10.18653/v1/2020.acl-main.417

Brook, J. W. (2022). ARW Mini Thesis: Towards Improving Neural Machine Translation for Low-Resource Languages (Unpublished).

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the Role of Bleu in Machine Translation Research, 249–256. https://homepages.inf.ed.ac.uk/pkoehn/publications/bleu2006.pdf

Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2020). A survey of model compression and acceleration for deep neural networks.

Chollet, F. (2021). English-to-Spanish translation with a sequence-to-sequence Transformer. https://keras.io/examples/nlp/neural_machine_translation_with_transformer/

Conroy, K. M. (2008). Celtic initial consonant mutations - nghath and bhfuil? *Boston College*. https://dlib.bc.edu/islandora/object/bc-ir:102094

CSO. (2023). Irish Language and the Gaeltacht - CSO - Central Statistics Office. https://www.cso.ie/en/releasesandpublications/ep/p-cp10esil/p10esil/ilg/

Dowling, M., Lynn, T., Poncelas, A., & Way, A. (2018). SMT versus NMT: Preliminary comparisons for Irish. *Proceedings of the AMTA 2018 Workshop on*

Technologies for MT of Low Resource Languages (LoResMT 2018), 12–20. https://aclanthology.org/W18-2202

Fadaee, M., Bisazza, A., & Monz, C. (2017). Data Augmentation for Low-Resource Neural Machine Translation. https://doi.org/10.18653/v1/p17-2090

Hajlaoui, N., Kolovratn'ik, D., Väyrynen, J. J., Steinberger, R., & Varga, D. (2014). DCEP -Digital Corpus of the European Parliament, 3164–3171. http://www.lrec-conf.org/proceedings/lrec2014/pdf/943_Paper.pdf

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.560

Kingma, D., & Ba, J. L. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://arxiv.org/abs/1412.6980

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Lankford, S., Afli, H., & Way, A. (2022). Human Evaluation of English–Irish Transformer-Based NMT. *Information*, *13*(7), 309. https://doi.org/10.3390/info13070309

Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *ScienceDirect*, *3*, 71–90. https://doi.org/10.1016/j.aiopen.2022.03.001

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2022). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2211.02001

Lynn, T. (2022). Report on the Irish Language. *European Language Equality*. https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_20__Language_Report_Irish_.pdf

Mac Giolla Chríost, D. (2004). *The Irish Language in Ireland: From Goídel to Globalisation*. http://ci.nii.ac.jp/ncid/BA7064182X

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. *arXiv (Cornell University)*. https://arxiv.org/pdf/2006.07264.pdf

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/pdf/1301.3781

nlp.Irish. (2020). https://nlp.irish/

OpenAI. (2023). ChatGPT. https://chat.openai.com/

Papineni, K., Roukos, S., Ward, T. J., & Zhu, W.-J. (2002). BLEU. *Meeting of the Association for Computational Linguistics*. https://doi.org/10.3115/1073083.1073135

Pexels. (2021). White Cat on Piano · Free Stock Photo. https://www.pexels.com/photo/white-cat-on-piano-6853287/

POS Tagging. (2019). https://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art)

Quteineh, H., Samothrakis, S., & Sutcliffe, R. (2020). Textual data augmentation for efficient active learning on tiny datasets. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7400–7410. https://doi.org/10.18653/v1/2020.emnlp-main.600

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., & Kaur, R. (2021). Neural Machine Translation for Low-resource Languages: A Survey. *ACM Computing Surveys*, *55*(11), 1–37. https://doi.org/10.1145/3567592

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*. https://arxiv.org/abs/1609.04747

Settles, B. (2011). *Active Learning*. Morgan & Claypool Publishers.

Stanojević, M., & Sima'an, K. (2014). BEER: BEtter Evaluation as Ranking. *Workshop on Statistical Machine Translation*. https://doi.org/10.3115/v1/w14-3354

StatMT. (2018). Parallel Corpus Filtering Task - EMNLP 2018 Third Conference on Machine Translation. https://www.statmt.org/wmt18/parallel-corpus-filtering.html

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schluter, P. (2012). DGT-TM: A freely available Translation Memory in 22 languages, 454–459. http://www.lrec-conf.org/proceedings/lrec2012/pdf/814_Paper.pdf

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. https://doi.org/10.18653/v1/p19-1355

Tan, L., Dehdari, J., & Van Genabith, J. (2015). An Awkward Disparity between BLEU / RIBES Scores and Human Judgements in Machine Translation. *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, 74–81.

Tatoeba: Collection of sentences and translations. (2006). https://tatoeba.org/en/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Neural Information Processing Systems*, *30*, 5998–6008. https://arxiv.org/pdf/1706.03762v5

Wang, R., Marie, B., Utiyama, M., & Sumita, E. (2018). NICT's Corpus Filtering Systems for the WMT18 Parallel Corpus Filtering Task. https://doi.org/10.18653/v1/w18-6489

Wang, X., Pham, H., Dai, Z., & Neubig, G. (2018). SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 856–861. https://doi.org/10.18653/v1/D18-1100

Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, U., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation.