Is Neural Machine Translation viable for Low-Resource Languages? An experimental study of the Irish Language

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Jack Quigley

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Dr. Ted Pedersen

July 2025

Dedication

To my family, friends, advisor, and professors whose support and encouragement made this research possible.

Abstract

Transformer-based Neural Machine Translation (NMT) models are Large Language Models (LLMs) designed and developed for translating between two or more given languages. These are typically most successful in the context of high-resource languages, languages with plentiful amounts of available online text corpora, such as English, Spanish, or French. In contrast, languages with limited corpora are known as low-resource languages and tend to be overlooked or underrepresented, like Basque, Pashto, or Ojibwe. One of these low-resource languages is Irish (Gaeilge), which has approximately 1.9 million total speakers as of 2022, and an extremely limited pool of publicly available datasets and machine translation systems. In response to this shortage, we created three bilingual English-Irish datasets and three transformer models for translating from English to Irish. Our models were then evaluated on four automatic evaluation metrics, BLEU, TER, CHRF, and METEOR, and demonstrated promising results across all our datasets.

Contents

List of Tables

List of Figures

1 Introduction

In this research, we attempted to answer whether modern Neural Machine Translation (NMT) systems are viable tools for translating low-resource languages, specifically for the Irish language. Modern NLP research focuses on only approximately 20 of the world's 7,000+ languages, leaving a significant number of languages unexplored, which can be classified as low-resource languages (Magueresse, Carles, and Heetderks 2020). These are languages that can be classified as "less studied, resource scarce, less computerized, less privileged, less commonly taught, or low density, among other denominations." One of these low-resource and understudied languages in NLP is the Irish language, which faces many challenges in the digital era compared to other dominant languages, such as English. In this study, we worked to answer the following research questions:

1. How competitive is fine-tuning a pre-trained Transformer-based neural machine translation (NMT) model compared with other machine translation systems for translating between English and a low-resource language, such as Irish?

2. How can the amount and quality of training data affect the performance of Transformer-based neural machine translation (NMT) models on BLEU, TER, CHRF, and METEOR evaluation scores for translating between English and a low-resource language, such as Irish?

We hypothesize that transformer-based Neural Machine Translation (NMT) models are a viable and competitive tool for translating between English and a low-resource language, such as Irish. Additionally, we hypothesize that a larger amount of low-quality translation training data will have less of an effect on the performance of fine-tuned NMT models compared to smaller high-quality translation training data.

We created three English to Irish bilingual datasets, a Legal, Web, and Combined version. These datasets were used to fine-tune three separate English to Irish NMT transformer models, E2G-Legal (E2G-L), E2G-Web (E2G-W), and E2G-Combined (E2G-C). Our models were then evaluated using the BLEU, TER, CHRF, and METEOR automatic evaluation metrics, on each of our three datasets, and averaged. We found that our E2G-L model is less competitive overall, while in contrast, our E2G-W and E2G-C models remain competitive and show improvements across all metrics. In the following section, we describe the state of the Irish language and our motivations for creating an English to Irish translation system.

1.1 The Irish Language

The Irish language is recognized as the first official language of Ireland. However, it was not until 2007 that the Irish language became an official part of the European Union (EU), and it was not until January 1, 2022 that the language was granted full official status, making Irish the 23rd official language of the EU (European Commission 2022). This is significant because it ensures that more EU resources and documents will be translated into Irish, just as other official EU languages, and that future resources will be translated, resulting in more Irish data becoming available.

According to Bradley 2014, in 1881, nearly one million people out of a population

of 3.87 million claimed they were "able to speak Irish," which then drastically declined to around 553,000 people, of a population of 3.13 million in 1911 (Government of Ireland 2012)[1]. However, a century later, in 2011, the number of speakers nearly tripled to around 1.7 million speakers of a population of 4.8 million. This implies that while the population was expanding, the number of speakers was declining, or, potentially, people were no longer learning the language to the same degree. However, Bradley 2014 also mentions that the census had failed to distinguish between people who speak Irish as their first versus second language, or what the definition of being "able to speak" Irish means. This raises another concerning implication that the number of speakers may be fewer than the census claims, indicating the language could be at greater risk than previously assumed.

According to the 2022 Ireland Census, there was a 6% increase in the number of people stating that they could speak Irish, totaling 1,873,997 in 2022 compared to the 1,761,420 in 2016 (*Census 2022* 2022). The census shows a growing number of speakers, but as stated earlier, we cannot be fully confident that the number of speakers is growing as much as they state, and there is still a likely need for the promotion and preservation of the Irish language.

1.2 Benefits of an Irish Translation System

In this section, we discuss some of our motivations for creating an Irish NMT system for translating from English to Irish, along with reasons why it is important. Due to a lack of digital resources for its accessibility and preservation, the Irish language faces many challenges in the digital era compared to other dominant languages, such

[1]All total population statistics of Ireland were taken from previous work by Government of Ireland 2012.

as the most prominent, English. Currently, there are no publicly released NMT systems focusing on supporting English to Irish translations, and English-Irish bilingual datasets are scarce. There are available systems that translate between languages (including Irish), but we failed to identify any machine translation systems primarily focused on translating into Irish. Additionally, opportunities to incorporate Irish into digital platforms, educational systems, and global exchanges will be severely limited without a reliable English to Irish translation system, as everything will need to be manually translated by individuals, resulting in higher costs and increased time for adding related Irish translations for new and existing digital platforms. Thus, filling these gaps will enable accurate and context-aware translations, promoting linguistic equity and ensuring the continued survival and expansion of the Irish language.

Work by Ní Chiaráin et al. 2023 investigated the potential for LLMs to be utilized as language tutors and as tools for beginner language learners to start composing stories without constant direction from teachers outlining specific tasks for them to write about. However, they found that ChatGPT performed worse than human tutors, but stated that language tutoring may be a domain where LLMs could be more effectively utilized and improved upon. One improvement we believe could help improve LLMs for language tutoring is using a more specialized model for translating between the respective languages, to help avoid teaching or using incorrect translations. A previous study by Dowling, Lynn, et al. 2018 showed how an out-of-the-box NMT system had achieved a lower translation quality when compared to a specifically tailored Statistical Machine Translation (SMT) system. This raises the question of how much better-quality translations a modern transformer model, pre-trained on multilingual data including Irish, produces when fine-tuned specifically for English-Irish translations?

A study by Duh et al. 2020 had shown that NMT systems can be competitive with related SMT systems for low-resource African languages, but only with sufficient hyperparameter tuning. Their findings also state that NMTs have the potential to benefit more from including additional data than SMTs. Additional work by Lankford, Afli, and Way 2024 showed the importance of hyperparameters in training in the context of Irish. In these studies, the researchers focus more on the models and the effects of hyperparameter tuning. However, they do not explore the extent to which selecting a dataset can influence the results. This leads to our final motivation for creating an Irish NMT system, which is to explore the extent different training data can affect the performance of the model's translations.

2 Related Work

In this chapter, we investigate the possibility of developing an English to Irish Neural Machine Translation (NMT) system to help promote and preserve the Irish language, by looking into previous translation systems for the Irish language. The Irish language, or Gaeilge, holds great cultural and historical significance in Ireland. It is recognized as Ireland's national and first official language, emphasizing its significance in the country's identity and history. However, despite its official status, the language faces challenges in terms of daily usage, including a decreasing population of native speakers and inadequate practical applications in many areas of modern digital life. According to Li, Brugha, and Gallagher 2017, in 2006, around 1.6 million people, out of a total population of 4.2 million, reported to have a communicational ability in Irish. However, only 84,000 of the 1.6 million claimed daily usage, with 31,000 being "school-going aged," implying they predominantly utilized the language for academic purposes (Li, Brugha, and Gallagher 2017).

Previous research efforts focused on developing translation systems for the Irish language have predominantly employed either, or a combination, of Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) models. However, there is still a notable lack of publicly available research on these approaches regarding the Irish language and the effectiveness of these approaches in translating between Irish and English.

2.1 Statistical Machine Translation (2003-2016)

Statistical Machine Translation (SMT) is a group of machine translation methods that leverage statistical methods to model the translation process, typically relying on large bilingual text corpora and probability algorithms. These systems aim to learn how to translate by identifying patterns and relationships between source and target languages in their training data. The most prominent SMT approaches include phrase-based translation, syntax-based translation, and hierarchical phrase-based translation. Phrase-based translation involves taking a source sentence and first splitting it into phrases, then translating phrase-by-phrase with some reordering of phrases in the target sentence, resulting in more context-aware and accurate translations (Koehn, Och, and Marcu 2003). Syntax-based translation takes advantage of syntactic structures from the source and target languages, making it suitable for more linguistically complicated texts (Koehn 2009). Hierarchical phrase-based translation expands on phrase-based models by including hierarchical rules that allow the translation of nested structures, bridging the gap between phrase-based and syntax-based approaches (Chiang 2005). These methods, which are not all-inclusive, collectively display the applications of SMT, each addressing unique issues in delivering accurate and fluid translations with statistical approaches. The most prominent SMT models for the Irish language are Tapadóir (Dowling, Cassidy, et al. 2015) and IRIS (Arcan et al. 2016).

Tapadóir (2015)

Previous work by Dowling, Cassidy, et al. 2015 created an SMT model, Tapadóir, built using the Moses toolkit (Koehn, Hoang, et al. 2007). This is an open-source toolkit for SMT with an additional motivation of extending phrase-based transla-

tions, thus classifying Tapadóir as an Irish-English phrase-based SMT model. They were able to show that Tapadóir was able to outperform Google Translate with an increased BLEU score of +10 points during their tests on domain-specific test data. However, their model failed to outperform Google Translate on TER with a lower score of -3 points.

IRIS (2016)

Previous work by Arcan et al. 2016 utilized similar SMT methods to create their model, IRIS. Similar to Tapadóir, IRIS is a phrase-based SMT framework built using the Moses toolkit to generate their translation models. In their evaluations, they found that IRIS was able to outperform Google Translate in both language directions, English-Irish and Irish-English, on all three of their evaluation metrics: BLEU, METEOR, and chrF. However, despite their models showing improvements as more data was added to IRIS, manual evaluators found that translation quality remained low.

2.2 Neural Machine Translation (2015-2024)

Neural Machine Translation (NMT) is the current cutting-edge approach to machine translation that utilizes neural networks for the translation process. Unlike SMT methods, NMT systems are more data-driven and learn from complex relationships and patterns between languages directly from large bilingual corpora. Several key architectures have shaped the development of NMT, including: Feed-forward neural networks (FNNs) (Bebis and Georgiopoulos 1994), Recurrent Neural Networks (RNNs) (J. Chung et al. 2015), Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber 1997), and most recently Transformer-based models

(Vaswani et al. 2017).

Transformer models, introduced by Vaswani et al. 2017, revolutionized NMT by making use of self-attention mechanisms to capture global dependencies in sequences, and since their introduction, most modern NMT approaches have been built on this architecture or similarly constructed architectures.

GaBERT (2022)

Previous work by Barry et al. 2022 developed a model based on the BERT architecture (Devlin et al. 2019), named gaBERT, that focused on Dependency Parsing, Cloze testing, and Multiword expression (MWE) Identification tasks. The gaBERT model was found to show improvements compared to three baseline models: mBERT (Devlin et al. 2019), WikiBERT-ga (Pyysalo et al. 2020), and XML-R_{BASE} (Conneau et al. 2020). Their model was solely trained and made for Irish language tasks, but did not include any machine translation tasks between Irish and English or other languages.

UCCIX (2024)

Previous work by Tran, O'Sullivan, and Nguyen 2024 developed the Irish-eXcellence LLM (UCCIX), based on the Llama 2-13B architecture, which, in the context of English-Irish machine translation, was evaluated using BLEU-4 on the gaHealth dataset developed by Lankford, Afli, and Way 2021 for English to Irish and Irish to English. Although this study did conduct MT training between English and Irish, it was not a primary focus as they also conducted Cloze testing and training on an Irish Question Answering task. Their study found that their baseline models had some understanding of Irish text and could translate from Irish to English but failed

to generate coherent Irish translations (Tran, O'Sullivan, and Nguyen 2024).

Lankford, et al. (2024)

Additionally, research by Lankford, Afli, and Way 2024 compared Transformer models with RNN models and found their Transformer models to outperform all compared RNN models. Their study showed that hyperparameter optimization for their Transformer model was crucial, resulting in an improvement in the BLEU score of +7.8 points compared to their baseline RNN model. When compared against Google Translate, they found their model outperformed on BLEU with an increased score of +14.2. However, their model was found to under-perform on TER with a decreased score of -11. However, they did not discuss releasing their model to the public or specify whether they used the baseline transformer model, introduced by Vaswani et al. 2017, or if they modified the baseline model.

3 Consulting Irish Organizations

This section describes how we contacted Irish organizations to collect stakeholder feedback on the use of LLMs and how they could be best used to aid in the preservation, promotion, and/or education of the Irish language. We refer to "Irish organizations" as institutions or companies based in Ireland that work with the language, either in preservation, promotion, or educational and academic contexts. A list of the contacted organizations and the outcomes of those interactions are listed in A.1.

3.1 Contacting Organizations

Our process for finding and reaching out to an organization was based on several factors. To find our initial list of possible organizations, we looked through various online sources and conducted basic Google searches. We found most organizations via gaeilge.ie[1] and peig.ie[2], where both sites provided a list of Official Irish bodies and/or leading Irish language organizations. Then, our first requirement was that the organization had to work with the language in some capacity, whether in education or academia (teaching the language or conducting research), in promoting the language, preserving the language, or working with native speakers. The second requirement was that the organization had a publicly viewable contact page for the relevant group working with the language or speakers. Finally, organizations predominantly focused

[1] https://www.gaeilge.ie/?lang=en
[2] https://peig.ie/en/organisations/

on running Irish-related events or producing broad Irish media were also excluded. This is primarily because they don't work directly with speakers of the language and are more focused on event planning and coordinating.

Our primary reason for contacting native Irish organizations is related to similar concerns raised by Caballar 2023 about Indigenous languages. In their article, they discuss concerns for data privacy and ensuring that technologies that utilize under-resourced languages remain in the control of the speakers. One primary concern raised is when governments, academic institutions, or other organizations collect data from Indigenous communities, they frequently have the power to deny access to their technologies or use them for external purposes without the approval of these communities. Therefore, it was in our best interest to create our datasets and language models to be utilized and expanded by future researchers, while avoiding any potential data privacy concerns.

3.2 Outcomes

During our interactions with these organizations, the primary shared desire was a universal translation system, so that individuals who wanted to continue speaking and using Irish could readily communicate with those who spoke, or preferred, an entirely different language. One way we can work towards this goal is to make translation systems to ensure Irish can compete with English and other major languages in the digital space.

One interview shared a copy of the Digital Plan for the Irish Language, a national project aimed at assisting the Irish language community in building, revitalizing, and modernizing the language within the technology sector. The plan focuses on technologies in the field of speech and language to keep up with other major European

languages in the digital age. The Digital Plan discusses some benefits of an MT system, including reducing translator workloads and improving productivity by being able to provide translators with powerful tools to help carry out their workload. Additionally, they point to numerous studies demonstrating how the incorporation of MT tools into existing workflows increases efficiency in academic and industry research. They discuss that when working on novel texts, where translation memory matches may be low, translators have higher productivity and translation quality when using MT tools (Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media 2022). The plan brings up several concerns in regards to creating NMT systems for Irish. Most importantly, the plan states how it is crucial for the future research and development of MT technologies for Irish that it does not become dependent on global technology companies.

Another concern raised is the need for human evaluators for reliably judging the usefulness of MT translation outputs. While automatic metrics are useful guidelines, they are not sufficiently robust indicators of the usefulness of MT outputs in professional environments (Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media 2022). Applications for MT discussed in the plan mention how these systems can help counteract English-dominated online platforms by helping translate them into Irish, allowing for Irish speaking communities to arise online. Additionally, it's mentioned how speech and language technologies can enable governments and public institutions to provide opportunities for all individuals to interact in Irish.

A separate concern brought up in both meetings was the lack of an official version of the language and how there's three distinct versions of the language. This was initially brought up as a concern for writing books for children and how choosing one version of the language can isolate certain groups because there are some parents

that don't want their children learning a different version of the language from them. According to Hickey 2023, the different dialects, or dialect areas—South, West, and North—result from a broader geographical distribution of Irish throughout the country. Southern Irish refers to areas in the southern province of Munster, Western Irish refers to forms of Irish spoken to the west of Galway and on the Aran Islands, and Northern Irish refers to Irish spoken in areas of West Ulster.

Another topic discussed was the encroachment of English into the daily lives of Irish speakers. More specifically, it was discussed how even in the deepest Gaeltacht regions, most phones are introduced in English by default and many Irish speakers keep their phones in English either because it doesn't work very well in Irish or because they don't know how, or it is not very simple, to change their phones into Irish. This is in addition to many children sitting at home using a popular streaming service or other social media platforms, which are predominantly in English. Many Irish schools primarily teach in English, offering Irish classes, except for Irish-medium schools.

Finally, it was also discussed how comparable LLMs are to the introduction of the printing press. It was discussed how when the printing press was introduced, it enabled some languages, especially those who adopted the technology early, while those who didn't embrace the technology struggled afterward to adopt and utilize it meaningfully.

4 Data

In the development of our datasets, we leveraged publicly available existing resources to create three distinct bilingual datasets tailored for machine translation between English and Irish. While searching for publicly available datasets, we found that a majority of them fell into one of two categories: Legislative or Web-crawled. For each selected dataset, we use the available English-Irish parallel corpora.

4.1 Legislative datasets

Legislative datasets were sourced from either the government of Ireland, the European Union, or other similar governmental bodies. Legislative datasets can be classified as official government documents established within the legal domain, including legal sessions, press releases, and legal documents. These can be considered high-quality bilingual datasets due to the severity that poor translations could cause, such as Irish speakers misunderstanding legal documents. Therefore, we assume that these datasets are human-translated, or at least human-reviewed. These datasets include:

DCEP (Digital Corpus of the European Parliament): Parallel multilingual corpus comprising the majority of documents published on the official European Parliament website, covering a wide range of topics, including press releases, session records, and legislative documents produced between 2001 and 2012

(Hajlaoui et al. 2014)[1].

DGT-TM (DGT-Translation Memory): Parallel multilingual corpus derived from European Union legislative texts in 24 official EU languages provided by the Directorate-General of Translation and taken from one of the European Commission's shared translation memory within EURAMIS (European Advanced Multilingual Information System) (Steinberger et al. 2013)[2]. However, due to the complexity and volume of files in the original data source, we opted to utilize an alternative public version of the dataset available on Kaggle.com[3].

Gaois (Parallel English-Irish Corpus of Legislation): Parallel Irish-English corpus of aligned text segments derived from Irish legislative documents[4].

4.2 Web-crawled datasets

Web-crawled datasets were sourced from various online sources, either specially curated or from conducting web scraping to collect their data. Web-crawled datasets can be classified as unofficial datasets available online without a specific domain focus. The quality of these datasets is uncertain, as it is unclear whether they contain properly human-translated references or if they are the result of other machine translation methods. Consequently, we recognize that some translations may contain inaccuracies as well as the potential for these datasets to contain duplicate entries. These datasets include:

[1]More information regarding DCEP can be found at the following website: https://joint-research-centre.ec.europa.eu/language-technology-resources/dcep-digital-corpus-european-parliament_en

[2]More information regarding DCEP can be found at the following website: https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memory_en#conditions

[3]https://tinyurl.com/28uunm2x

[4]https://www.gaois.ie/en/corpora/parallel/data

ParaCrawl v9: Large-scale open-sourced parallel corpus built by crawling and aligning multilingual web content extracted, cleaned, and aligned from publicly accessible websites (Bañón et al. 2020).

HPLT (High Performance Language Technologies) language resources v1.2: Large English-centric multilingual corpus obtained from CommonCrawl and under-used web crawls from the Internet Archive covering 75 monolingual and 18 parallel corpora. (Gibert et al. 2024).

ELRC: Parallel bilingual corpus of texts crawled from multilingual websites and created within the framework of the European Language Resource Coordination (ELRC) released by the EU Directorate-General for Communications Networks, Content and Technology (European Commission, Directorate-General for Communications Networks, Content and Technology 2019)[5].

4.3 Preprocessing

One concern for the given datasets was the potential for them to contain problematic bias that could be inherited from the data source. Vallor 2024 discusses the potential for AI to change our values, for better or worse, and can result in our virtues being harder to recognize in ourselves and others, which can lead to a deterioration of what we think a better character may be. Therefore, it is in our best interest to avoid any potential bias' where possible to ensure if our model makes an incorrect translation that it is not an obscene or subtly incorrect translation that could cause harm or confusion in any way. Additionally, it has been shown that LLMs have the

[5] https://commission.europa.eu/about/departments-and-executive-agencies/communications-networks-content-and-technology

potential to generate undesirable output, known as hallucinations, that are nonsensical or unfaithful to the provided input (Ji et al. 2023). By removing translations that may contain unwanted content in our datasets can help reduce the potential for our trained model to generate obscene or unwanted hallucinations.

Before conducting preprocessing on all of our datasets and preemptively glancing at their content, it was clear that some specific preprocessing had to be done. The most significant was the HPLT dataset. We noticed a considerable amount of explicit content, which was found to be about 26% of the original dataset. Due to the large percentage of explicit content in the dataset, we went through the dataset and removed translations that contained explicit content. Explicit content detection was conducted using the better-profanity Python library[6]. An example of how we used better-profanity to conduct our processing of a pandas dataframe of HPLT can be seen in appendix B.1. Explicit content removal was only performed on the HPLT dataset since there was no obvious sign that other datasets contained a large percentage of this type of content.

Each dataset was individually preprocessed using regular expressions (regex) to remove any leading newline characters or leading digits. This was done to avoid unnecessary leading characters that do not add relevant information, such as bullet points or line numbering that are not necessarily applicable outside the original list. The specific code we used can be found in appendix B.1. Notably, there is a chance for unclosed parentheses or other similar symbols to appear because of how we remove any starting symbols. This is unlikely to occur because of the likelihood of sentences starting with parentheses or similar symbols, but it is a potential concern. Examples of our regex code and how it can affect our data can be found in appendix B.

After preprocessing all individual datasets, they were concatenated into their re-

[6] https://pypi.org/project/better-profanity/

spective master datasets: Legislative (Legal), Web-crawled (Web), and a combination of them (Combined). Each master dataset was then converted to Pandas dataframes for further processing by removing rows containing a NULL value in either language and eliminating duplicates from each language column. This step ensured that every column had a respective translation and verified that any repeated English translations had identical Irish translations, and vice versa. Our example code can be found in appendix B.1. Furthermore, once the Legal and Web datasets were combined and preprocessed, they were joined together to form the Combined dataset. Once combined, another round of duplicate and null removals was conducted to ensure that any content from the Legislative and Web datasets had no overlapping duplicates.

The final preprocessing step involved dividing each master dataset into three respective splits for training, testing, and validation/evaluation purposes. Specifically, 80% of each dataset was assigned for training, 10% for testing, and another 10% for validation. The sizes of each dataset and their splits can be found in table 4.1.

Dataset	Split	Number of Translations
Legal	Train	478,729
Legal	Test	59,842
Legal	Valid	59,841
Legal	Total	598,412
Web	Train	2,400,472
Web	Test	300,059
Web	Valid	300,059
Web	Total	3,000,590
Combined	Train	2,821,645
Combined	Test	352,706
Combined	Valid	352,706
Combined	Total	3,527,057

Table 4.1: Overview of datasets, including train/test/valid splits, the number of English-Irish translations in each subset, and the total number of entries across all splits.

5 Model

This chapter presents our methodology for building and evaluating our English to Irish MT models. We discuss our selection of a pre-trained base model, considering language coverage, model architecture, and potential environmental impacts. We then describe our fine-tuning procedure, including the training configuration and relevant hyperparameters used for training three English-Irish MT models: E2G-Legal (E2G-L), E2G-Web (E2G-W), and E2G-Combined (E2G-C). Finally, we describe our methodology for evaluating our models and the automatic metrics used to assess the translation quality of our models.

5.1 Base Model

When considering our selection of a base model for fine-tuning, we noticed that many modern MT models are built on the original transformer model by Vaswani et al. 2017. Additionally, instead of the typical process of pre-training and fine-tuning a base Transformer, which requires significant amounts of time and energy, we chose to utilize transfer learning. This method enables us to start with a previously pre-trained model, avoiding a random initialization of parameters and leading to a more efficient training, even with smaller amounts of data (Magueresse, Carles, and Heetderks 2020). Additionally, previous work by Chi et al. 2024 found that a pre-trained English language model can learn both language-specific and language-agnostic information beneficial for non-English models in low-resource scenarios. Therefore, there's

potential for models trained on more languages to benefit from both language-specific and language-agnostic information contained within those languages, implying benefits for choosing a model with more language representations.

We subsequently proceeded to select from popular MT models promoted by HuggingFace's translation fine-tuning tutorial[1]. We subsequently narrowed down the model options further based on existing research publications and, most importantly, the inclusino of some form of pre-training containing the Irish language. Consequently, our final model choices were the following:

- M2M100[2] as introduced by Fan et al. 2020. M2M100 is a fully multilingual MT model trained on direct translations between 100 languages and supports many-to-many translations without pivoting through English.

- mT5[3] as introduced by Xue et al. 2021. mT5 is a multilingual adaptation of the original T5 model by Raffel et al. 2023 and trained on the mC4 corpus, a massive dataset covering 101 languages.

- NLLB[4] as introduced by NLLB Team et al. 2022. NLLB, short for No Language Left Behind, is a multilingual model developed to improve translation quality for low-resource languages, breaking the 200 language barrier.

- UMT5[5] as introduced by H. W. Chung et al. 2023. uMT5, Universal mT5, is an extension of the mT5 model that incorporates task-specific prompts enabling a unified approach to multilingual and cross-lingual learning. Additionally, uMT5 uses a shared architecture for each task or language.

[1] https://huggingface.co/docs/transformers/en/tasks/translation
[2] https://huggingface.co/docs/transformers/en/model_doc/m2m_100
[3] https://huggingface.co/docs/transformers/en/model_doc/mt5
[4] https://huggingface.co/docs/transformers/en/model_doc/nllb
[5] https://huggingface.co/docs/transformers/main/en/model_doc/umt5

From these models, we worked to compare them and narrow down our selection to a single model. For our comparisons of these models, we looked at the smallest-sized model versions in terms of parameters: google/mt5-small (mT5), google/umt5-small (umT5), nllb-200-distilled-600M (NLLB), and m2m100_418M (M2M100). We compare the smallest model sizes in an effort to achieve smaller model sizes, which have been shown to reduce potential CO2 emissions without compromising the performance of our future models (Liu and Yin 2024). Additionally, research by Ogueji, Zhu, and Lin 2021 found that training models on smaller curated datasets for low-resource languages can be competitive compared to models with a larger number of parameters. Thus, it's important to prioritize smaller model sizes to increase model efficiency and contribute to a more sustainable MT environment.

We initially looked into the two models based on the T5 architecture (Raffel et al. 2023), mT5 and umT5. These models require prompts and are built to be used on multiple tasks (Xue et al. 2021; H. W. Chung et al. 2023). According to Xue et al. 2021, they found that their small and base mT5 models have the highest amount of accidental translations compared to their larger models. However, the larger umT5 and mT5 models have higher parameter counts in comparison to the base NLLB and M2M100 models. Thus, to use either of the T5 model variants in a way to avoid more accidental or hallucinated translations, we would have to give up a smaller parameter size model to get a similar performing base model. Furthermore, Liu and Yin 2024 found that the original T5 model produces much higher CO2 emissions compared to BERT-based models. Although they do not compare against the M2M100 or NLLB models, it was clear that there's a possibility that these T5-based models produce higher CO2 emissions.

During the investigation of the remaining two models, M2M100 and NLLB, we

observed that the configuration of the NLLB model was based on the same architectural configuration as the M2M100 model. The main difference between the two models is their parameter sizes: about 484 million for M2M100 and about 615 million for NLLB. Additionally, findings from Ogueji, Zhu, and Lin 2021 found that smaller-sized models can outperform others and be more reproducible despite having half as many parameters. Therefore, we concluded that it is in our best interest to fine-tune the M2M100 model due to its smaller parameter size, which as been shown to reduce our CO2 emissions as much as possible.

Ultimately, we chose to conduct fine-tuning on the M2M100 model on all three of our training datasets for English to Irish translation, subsequently naming our models E2G-Legal (E2G-L), E2G-Web (E2G-W), and E2G-Combined (E2G-C), standing for English to Gaeilge (Irish).

5.2 Fine-tuning Methodology

To setup our training, we first adapted the code from HuggingFace's translation task guide[6] for use with our datasets and base pre-trained model. We initialized our model and tokenizer from the "facebook/m2m100_418M" checkpoint, available on the HuggingFace Hub, using the M2M100ForConditionalGeneration and M2M100Tokenizer Python classes. We initialized our tokenizer with the src_lang (source language) set to "en" for English and the tgt_lang (target language) set to "ga" for Irish. Then, the respective training dataset for the desired model was loaded and tokenized using our tokenizer.

We implemented our training procedure using the Seq2SeqTrainer from Hug-

[6] https://huggingface.co/docs/transformers/en/tasks/translation#evaluate

gingFace's trainer module[7]. This was to utilize the trainer's support for distributed training on multiple GPUs and mixed precision for NVIDIA and AMD GPUs, in addition to its integration with the Hugging Face Hub to quickly upload the model after the completion of training.

We set our training arguments using the Seq2SeqTrainingArguments class, setting the number of training epochs to 3, optimizer to adafactor, learning rate to 2e-5, weight decay to 0.01, and load best model at end to true. We then initialized our Seq2SeqTrainer with our model, training arguments, training and test dataset loaders, tokenizer, and data collator (DataCollatorForSeq2Seq). We chose not to use any compute metrics method for our trainer to help speed up training and minimize the energy and resource consumption associated with calculating additional metrics.

Despite having access to NVIDIA A100 GPUs, we utilized a less expensive alternative by using an NVIDIA 3080Ti GPU for our training and evaluation processes. In comparison to an A100 GPU, it is a much more affordable and accessible GPU that allows for reproducible and comparable results that others can generate without needing access to expensive top of the line GPUs. Additionally, Liu and Yin 2024 discussed the trade-offs between the cost and accessibility of GPUs and training LLMs, discussing the importance of choosing faster GPUs and lower-parameter models to balance potential environmental and financial implications.

Ultimately, we fine-tuned three separate M2M100 models on each of our master datasets and named them, respectively: E2G-Legal (E2G-L), E2G-Web (E2G-W), and E2G-Combined (E2G-C). An example of our full training code used for our models can be found in B.2.

[7] https://huggingface.co/docs/transformers/v4.48.2/en/main_classes/trainer#transformers.Trainer

5.3 Evaluation Methodology

Similar to the previously discussed English-Irish MT models, we report our results based on several automatic evaluation metrics, including BLEU (Papineni et al. 2002), TER (Snover et al. 2006), CHRF (Popović 2015), and METEOR (Banerjee and Lavie 2005). We conducted evaluations separate from training utilizing HuggingFace's Evaluate Python library[8] to reduce initial training time and to allow for evaluations on each of the valid splits for our datasets.

We initialized the desired fine-tuned model for evaluation from the HuggingFace Hub, using half-precision weights to save GPU memory and increase speed, and then set the model to evaluation mode. Additionally, we utilize a CUDA device and use TensorFloat-32 (tf32) mode[9] for faster calculations at the risk of slightly less precise predictions. Then the tokenizer was initialized in the exact same way as we did in fine-tuning, using a M2M100Tokenizer from the "facebook/m2m100_418M" checkpoint and setting the source and target languages to English and Irish, respectively. For generating model predictions, the English input text was first tokenized with padding and truncation enabled. Then the model generated its prediction of an Irish translation with use_cache enabled, number of beams set to one, and a maximum length of 128. After the model generates a prediction, it is then de-tokenized for evaluation. Each of our fine-tuned models generated predictions for the "valid" split for each of our three datasets.

For evaluating model predictions, we first loaded each of our evaluation metrics using HuggingFace's evaluation library. Then we computed the metric scores by passing in the predicted translations and the reference Irish text which was then

[8] https://huggingface.co/docs/evaluate/en/index

[9] https://huggingface.co/docs/diffusers/en/optimization/fp16#tensorfloat\protect\discretionary{\char\hyphenchar\font}{}{}32

exported to a csv file with the metric and associated scores.

To demonstrate the process for calculating these metrics, we first examine a specific example for calculating the BLEU score with a candidate (predicted) translation and reference translation. A more in-depth description of the BLEU metric and how it works is discussed further in chapter 5.3.1.

Metric scores were calculated in their original range, either from 0.0 to 1.0 or 0.0 to 100.0. Accordingly, all metrics were then standardized to a range of 0.0 to 100.0 by multiplying by 100 for easier comparison of results. It's important to note that we cannot directly compare different measures, such as BLEU and CHRF, because they evaluate different aspects of the given translations. Outlined below are the evaluation metrics used for assessing model performance, along with a brief description of their purpose and an example of how they are calculated.

5.3.1 BLEU

The Bilingual Evaluation Understudy (BLEU) metric was derived from the widely successful word error rate metric used in the speech recognition community and was further modified to accommodate multiple reference translations and variations in word choice and word order (Papineni et al. 2002). The BLEU metric measures a modified unigram precision of a predicted translation compared to a reference translation. It does this by first finding the maximum frequency each word appears in the reference translation, then it takes the minimum number between that word's frequency and the maximum number of times that word appears in the reference translations, sums them, and divides them by the total number of words in the predicted translation. The final BLEU score also applies a sentence brevity penalty, penalizing translations whose lengths are too short. For our research, we use the

SacreBLEU implementation, a Python script designed to improve the reproducibility of the BLEU metric, for computing our BLEU scores (Post 2018). Comparative models evaluated on the BLEU metric include: IRIS (Arcan et al. 2016), UCCIX (Tran, O'Sullivan, and Nguyen 2024), Tapadóir (Dowling, Cassidy, et al. 2015), and the models from Lankford, Afli, and Way 2024.

BLEU Example: A candidate translation: "The the the the the the." is compared with a reference translation: "The food on the plate." Comparing using unigrams, we find that "the" appears six times in the candidate, two in the reference, and none of the other words in the candidate match the reference. Thus, the candidate translation has a modified unigram precision (P) of 2/6, or $\approx 0.\overline{33}$. Then, BLEU would calculate a sentence brevity penalty (BP) if the length of the candidate is larger than the reference. Thus, for this example, our candidate's length (6) is larger than our reference's length (5), setting BP to 1. Accordingly, we can compute the final BLEU score using the equation:

$$\text{BLEU} = \text{BP} \cdot exp(\ln P) = 1 \cdot exp(\ln 0.\overline{33}) = 0.\overline{33} = 33.3$$

5.3.2 TER

The Translation Edit Rate (TER) metric is defined as the minimum number of edits required to change a potential translation to exactly match one of the given references, then normalized by the average length of references. Essentially, the TER metric measures the number of edits a human translator would need to make to correct a machine-produced translation, divided by the average number of reference words. An edit can refer to insertions, deletions, substitutions of individual words, or shifts in word sequences (Snover et al. 2006; Post 2018). Therefore, the more edits

needed to correct a translation will lead to a higher TER score, indicating that a lower score represents a higher-quality translation. This is in contrast to the other metrics, where a higher score signifies a higher quality of translation. Comparative models evaluated on the TER metric include: Tapadóir (Dowling, Cassidy, et al. 2015) and the models from Lankford, Afli, and Way 2024.

TER Example:

A candidate translation: "The the the the the." is compared with a reference translation: "The food on the plate." Comparing the candidate and reference translations, we find that "the" appears twice in the reference but five times in the candidate, and the candidate is missing the words 'food', 'on', and 'plate'. Since "the" appears three more times than it should in the candidate and is missing three words from the reference, it would only require three substitution edits to fix the candidate to match the reference translation. Then, using the length of the reference translation of five, we can compute the final TER score using the equation:

$$\text{TER} = \frac{3}{5} = 0.6$$

5.3.3 CHRF

The CHRF metric is a character n-gram F-score designed to better reflect human judgments compared to BLEU and TER, inspired by other linguistically motivated F-scores based on Part-of-Speech (POS) tags and morphemes, thus similarly captures some morpho-syntactic features (Popović 2015; Popović 2017; Post 2018). CHRF is calculated by first finding the percentage of character n-grams in the prediction also present in the reference (chrP), then the percentage of character n-grams in the reference also present in the prediction (chrR). Then, the final CHRF score gets

calculated using the following equation:

$$\text{CHRF} = \frac{2 \cdot \text{chrP} \cdot \text{chrR}}{\text{chrP} + \text{chrR}}$$

Additionally, Popović 2015 describes CHRF as being language and tokenization-independent and has demonstrated strong correlations with human evaluations at both system and segment levels. Comparative models evaluated on the CHRF metric include: IRIS (Arcan et al. 2016) and the models from Lankford, Afli, and Way 2024.

CHRF Example:

A candidate translation: "The the the the the the." is compared with a reference translation: "The food on the plate." The first step is to count the number of characters in the candidate translation, which is 18, and in the reference translation, which is 17. Then we can find the number of characters from the candidate that match characters in the reference. To calculate this, the word "the" appears fully twice in the reference and one 'e' in the candidate matches with the 'e' in "plate" meaning there are a total of 7 matched characters. Next, we can compute chrP using the number of matched characters and the length of the candidate translation, and chrR using the number of matched characters and the length of the reference translation using the following equations:

$$\text{chrP} = \frac{7}{18} = 0.3889$$

$$\text{chrR} = \frac{7}{17} = 0.4118$$

Finally, we can compute the final CHRF score using the equation:

$$\text{CHRF} = \frac{2 \cdot 0.3889 \cdot 0.4118}{0.3889 + 0.4118} \approx 0.4$$

5.3.4 METEOR

The Metric for Evaluation of Translation with Explicit ORdering (METEOR) is an automatic metric for MT evaluation based on a generalized concept of unigram matching to compare a machine-produced translation with a human-provided reference translation (Banerjee and Lavie 2005). The METEOR score is based on the harmonic mean of precision and recall and calculated using a combination of unigram-precision, unigram-recall, and a measure of how well-ordered the matched words are in the MT compared to the reference translation. METEOR is calculated, first by finding the number of matched words (M) from the candidate translation to the reference translation and ordering them by exact match, stem match (e.g. read/reading), then synonym match (e.g. good/kind), where each word can only match once. Unigram precision (P) is calculated by dividing M by the total number of words in the candidate translation, and the unigram recall (R) is calculated by dividing M by the total number of words in the reference. Next, the Fmean is computed with a harmonic mean with most of the weight on recall using the following equation:

$$Fmean = \frac{10 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{R} + 9\mathrm{P}}$$

Then a penalty is calculated, to account for longer matches, by finding the fewest number of chunks, or groups of adjacent matched words that appear in the same order in both candidate and reference, divided by the number of unigram matches between the candidate and reference, which is cubed and then multiplied by 0.5. Accordingly, the final METEOR score is calculated via the following equation:

$$\mathrm{METEOR} = Fmean \cdot (1 - penalty)$$

METEOR was created to address weaknesses in BLEU and NIST (Doddington 2002), such as a lack of recall, use of higher order n-grams, lack of explicit word-matching between translation and reference, and the use of geometric averaging of n-grams. The only comparative model evaluated on the METEOR metric was the IRIS model (Arcan et al. 2016).

METEOR Example:

A candidate translation: "The the the the the the." is compared with a reference translation: "The food on the plate." Comparing the candidate and reference translaitons, we find that "the" appears six times in the candidate, two in the reference, and none of the other words in the candidate match the reference, meaning we have two matched words between the translations. Thus we can calculate P using the length of the candidate translation: $P = \frac{2}{6} \approx 0.333$ and R using the length of the reference translation: $R = \frac{2}{5} = 0.4$. Next, we can calculate our Fmean using the equation:

$$Fmean = \frac{10 \cdot 0.333 \cdot 0.4}{0.4 + 9(0.333)} = 0.3921$$

Then, we need to calculate the penalty with the fewest number of chunks (1) and the number of matched words (2) from the candidate and reference using the following equation:

$$\text{penalty} = 0.5 \cdot (\frac{1}{2})^3 = 0.0625$$

Finally, we can calculate our final METEOR score with the equation:

$$\text{METEOR} = 0.3921 \cdot (1 - 0.0625) = 0.3675$$

6 Results

We present the results of our evaluations on our three models, E2G-Legal (E2G-L), E2G-Web (E2G-W), and E2G-Combined (E2G-C). These models only differ in regards to what data they were fine-tuned on. Each model was evaluated on four automatic evaluation metrics, BLEU, TER, CHRF, and METEOR. We use four metrics to more effectively represent the performance of our models, as each metric evaluates a slightly different aspect of translation quality. This reduces any potential errors found from any metric, and using multiple metrics may help reduce bias or issues that any single metric may have. For each metric, each model was evaluated across three valid splits from our training datasets; Legal, Web, and Combined. We find that E2G-L on average underperformed both E2G-W and E2G-C on all metrics, whereas E2G-C outperforms both E2G-L and E2G-W models on all metrics across all datasets. E2G-W remained competitive with E2G-C on most metrics and datasets but failed to achieve the same results and consistency as E2G-C.

6.1 BLEU

All model scores on BLEU for each validation dataset split, and averaged, can be seen in figure 6.1. On average, E2G-L underperformed both E2G-W and E2G-C by a large margin of -18.5 and -22.1. However, E2G-L only underperformed by -4 and -6.3 on the Legal dataset in comparison to the best BLEU scores for the E2G-W and E2G-C models, respectively, with both E2G-Web and E2G-C getting

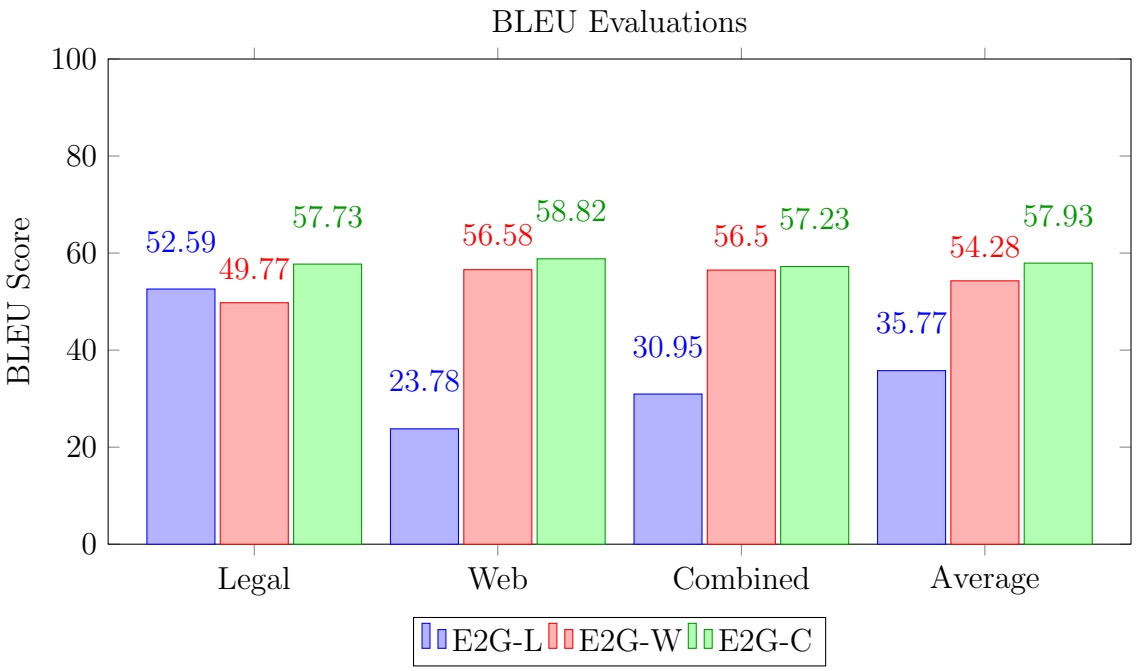

Figure 6.1: Comparison of model performances on the BLEU metric across Legal/Web/Combined datasets and the averaged score across all datasets.

their best BLEU scores on the web dataset. E2G-L appears to perform the best on the in-domain data compared to the other models, which were much less domain-specific in training. The E2G-W and E2G-C models are much more consistent across all datasets getting relatively similar scores. However, E2G-W appeared to slightly under-perform on the Legal dataset compared to other datasets. In contrast, E2G-C outperformed both of our other models on average by +3.6 and +18.5 across all datasets. Furthermore, E2G-C had the best performance across all models and datasets, with the highest score on the Web dataset at 58.82.

6.2 TER

All model scores on TER for each validation dataset split, and averaged, can be seen in figure 6.2. As mentioned in chapter 5.3.2, a lower TER score represents a

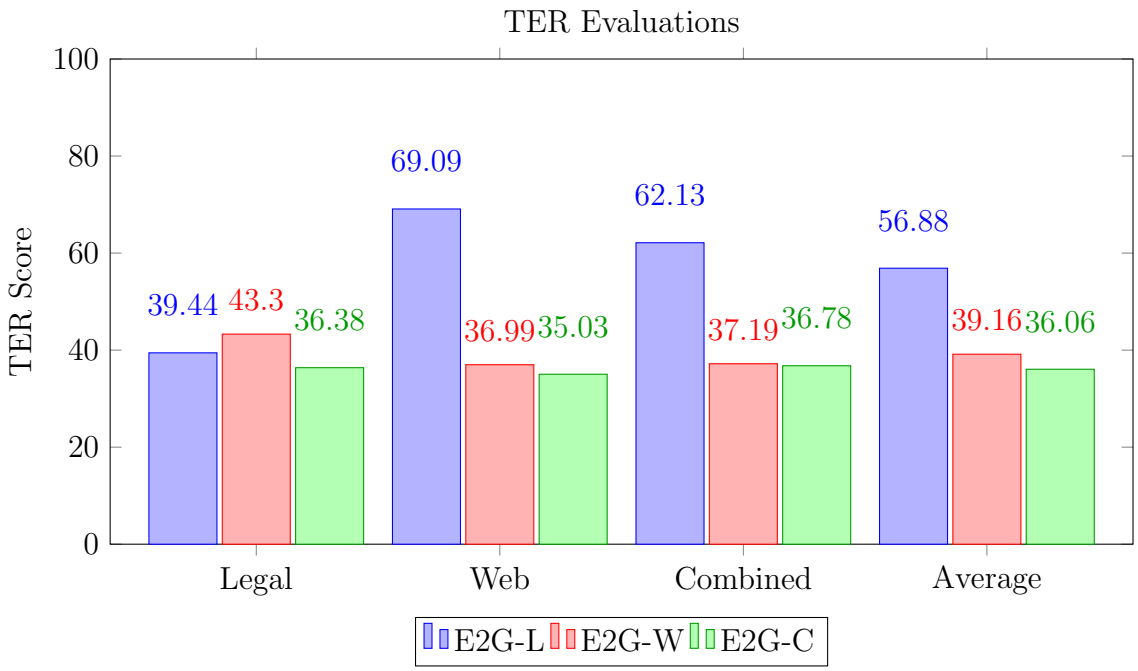

Figure 6.2: Comparison of model performances on the TER metric across Legal/Web/Combined datasets and the averaged score across all datasets.

better performance and fewer edits needed by a human translator. On average, E2G-L underperformed both E2G-W and E2G-C by a large margin of +17.72 and +20.82. E2G-L drastically underperformed E2G-W and E2G-C on both Web and Combined datasets. However, E2G-L had a better performance on the Legal dataset compared to E2G-W. E2G-L had the worst score at 69.09 on the Web dataset and the second worst score of 62.13 on the Combined dataset, scoring more than +25 on each. Both E2G-W and E2G-C were much more consistent across all datasets, but E2G-C had the best performances overall with an average TER score of 36.06 and best score on the Web dataset of 35.03. In contrast, E2G-W had the worst performance on the Legal dataset at 43.30, +3.86 compared to E2G-L and +6.92 against E2G-C.

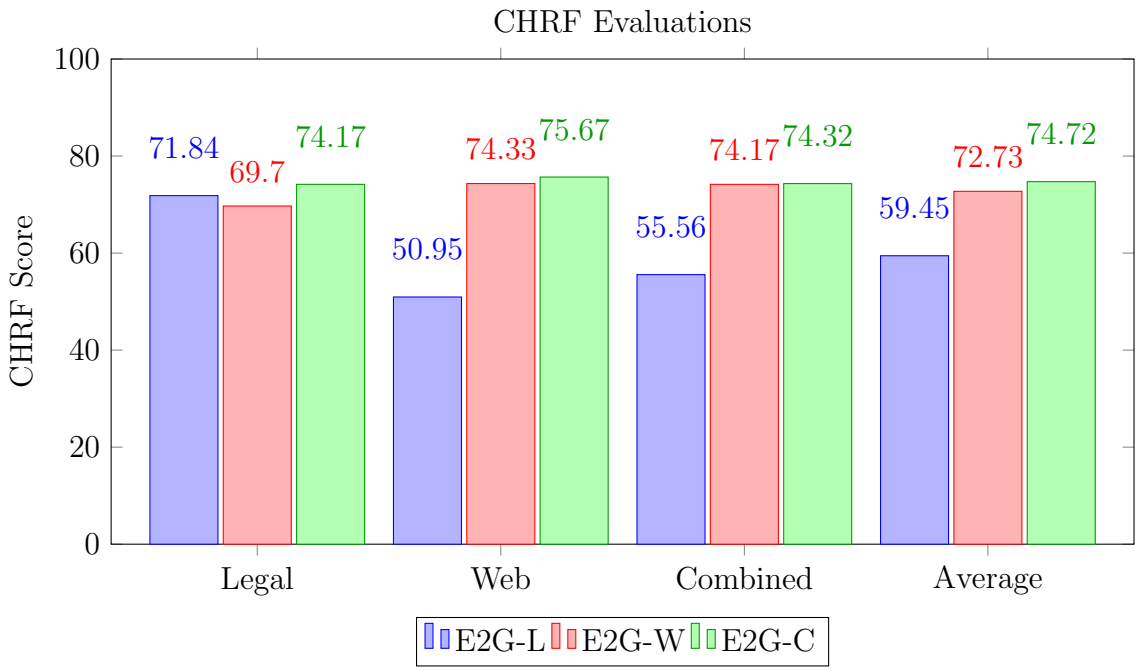

Figure 6.3: Comparison of model performances on the CHRF metric across Legal/Web/Combined datasets and the averaged score across all datasets.

6.3 CHRF

All model scores on CHRF for each validation dataset split, and averaged, can be seen in figure 6.3. On average, E2G-C outperformed E2G-L and E2G-W by +15.27 and +1.99 and showed the best performance on the Web dataset at 75.67. E2G-L had outperformed E2G-W on the Legal dataset by +2.14, but had drastically underperformed E2G-W and E2G-C on the Web and Combined datasets. E2G-L's average chrF score was the lowest at 59.45, while E2G-C had the best average score at 74.72.

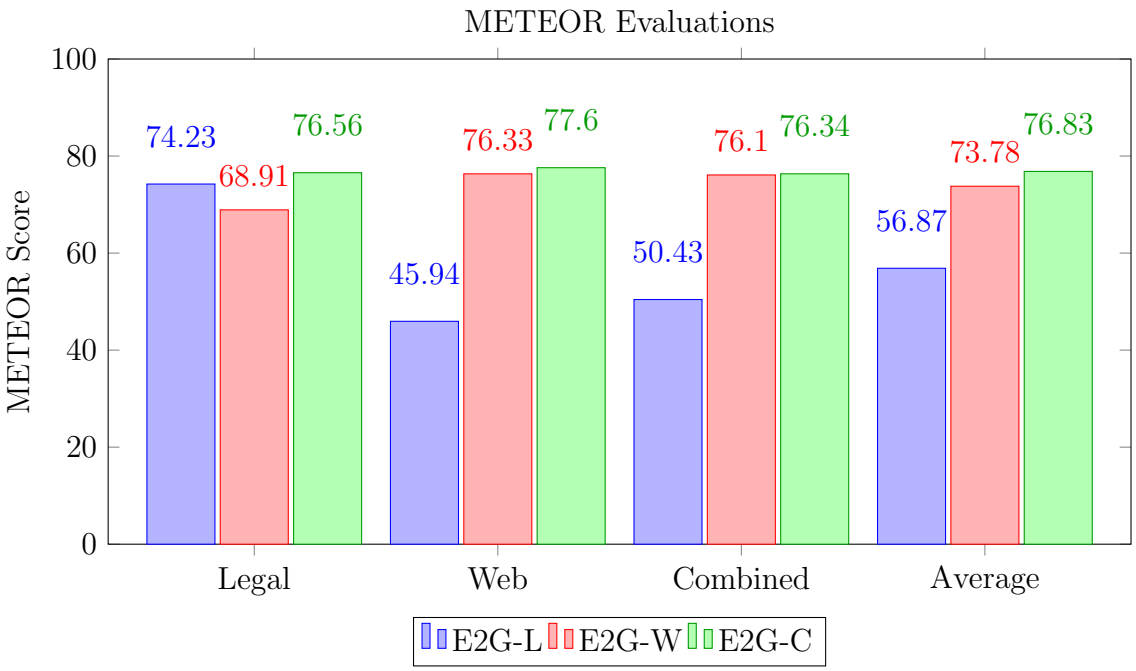

Figure 6.4: Comparison of model performances on the METEOR metric across Legal/Web/Combined datasets and the averaged score across all datasets.

6.4 METEOR

All model scores on METEOR for each validation dataset split, and averaged, can be seen in figure 6.4. E2G-C outperformed across all datasets, getting an average score of 76.83 and the best score on the Web dataset with a score of 77.6. E2G-C had the most consistent scores with a difference of -1.26 between it's highest and lowest scores. In contrast, overall E2G-L underperformed both other models on the Web and Combined datasets by more than -30 and -20, but outperformed E2G-W by +5.32 on the Legal dataset and only underperforming E2G-C by -2.33. On average, E2G-L underperformed E2G-W and E2G-C by -16.91 and -19.96 and had the overall worst score of 45.94 on the Web dataset.

6.5 Key Findings

Overall, we find that E2G-L on average underperformed both E2G-W and E2G-C on all metrics. Additionally, E2G-C outperformed E2G-W and E2G-L on all metrics across all datasets. E2G-W remained competitive with E2G-C on most metrics and datasets but did not show as consistent results and underperformed E2G-L on several dataset/metric pairs. Results for all of our model evaluations for each dataset and evaluation metric can be found in table 6.1.

Model	Dataset	BLEU	TER	CHRF	METEOR
E2G-L	Legal	**52.59**	**39.44**	**71.84**	**74.23**
E2G-L	Web	23.78	69.09	50.95	45.94
E2G-L	Combined	30.95	62.13	55.56	50.43
E2G-L	Average	35.77	56.88	59.45	56.87
E2G-W	Legal	49.77	43.30	69.70	68.91
E2G-W	Web	**56.58**	**36.99**	**74.33**	**76.33**
E2G-W	Combined	56.50	37.19	74.17	76.10
E2G-W	Average	54.28	39.16	72.73	73.78
E2G-C	Legal	57.73	36.38	74.17	76.56
E2G-C	Web	**58.82**	**35.03**	**75.67**	**77.60**
E2G-C	Combined	57.23	36.78	74.32	76.34
E2G-C	Average	**57.93**	**36.06**	**74.72**	**76.83**

Table 6.1: Evaluation results on all of our models, where the best scores for each metric and model combination are shown in bold, the best average for each metric is also in bold, and the best result for each metric is underlined.

7 Discussion

We analyze the results presented in chapter 6 and compare them against the previously discussed models in chapter 2. The models we compare to include the SMT models IRIS (Arcan et al. 2016) and Tapadóir (Dowling, Cassidy, et al. 2015), and the NMT models UCCIX (Tran, O'Sullivan, and Nguyen 2024) and the best-performing model presented by Lankford, Afli, and Way 2024, which was their "dgt-trans-bpe16k" transformer model, which we will be referring to as DGT-BPE16k. We compare these models to our three fine-tuned models, E2G-Legal (E2G-L), E2G-Web (E2G-W), and E2G-Combined (E2G-C). For our comparisons against other models, we take the highest published score for each metric from their respective studies. The evaluation metrics used were BLEU, TER, CHRF, and METEOR. BLEU measures the ratio of correctly predicted words to total predicted words. TER measures the needed edits to fix a machine-generated translation relative to reference words. CHRF evaluates the alignment of predicted and reference translations by balancing correct and missed words. Finally, METEOR measures the correctness and order of words of predicted and reference translations. Some of these metric scores were reported in the original range of 0.0-1.0. Therefore, where necessary, these scores were standardized to be in the range from 0.0-100.0, to make them comparable to our results.We will then compare those scores to the average metric scores, unless otherwise noteworthy, found from our models in chapter 6. Finally, it is important to note that the results of each model compared were evaluated using

different datasets and trained on differing training data. Therefore, our results may not fully reflect model performance, but they provide a general idea of how our models perform in comparison.

7.1 BLEU

The BLEU metric computes the number of correctly predicted words divided by the total number of predicted words. Thus, a higher BLEU score means there was a more precise number and accuracy of unigrams in the predicted translation compared to the reference translation(s). Comparative models evaluated on the BLEU metric include: IRIS (Arcan et al. 2016), UCCIX (Tran, O'Sullivan, and Nguyen 2024), Tapadóir (Dowling, Cassidy, et al. 2015), and DGT-BPE16k (Lankford, Afli, and Way 2024).

We found our E2G-L model underperformed Tapadóir by -7.31, IRIS by -20.0, and DGT-BPE16k by -24.73. However, E2G-L slightly outperformed UCCIX by +2.43. However, when compared with the E2G-L model's score on the Legal dataset, E2G-L had only slightly underperformed IRIS by -3.18 and DGT-BPE16k by -7.91 and outperformed Tapadóir by +9.51 and UCCIX by +19.25. Overall, E2G-L underperformed the compared models on BLEU, behind two of the models by a significant margin of over -20, and just marginally outperformed UCCIX. We found that E2G-L was much less competitive on BLEU compared to Tapadóir, IRIS, and DGT-BPE16k but was slightly more competitive compared to UCCIX, likely due to UCCIX not being a focused MT model but being trained for many tasks.

Our E2G-W model performed better than E2G-L, only underperformed compared to IRIS slightly by -1.49 and by -6.22 from DGT-BPE16k, and outperformed Tapadóir by +11.2 and UCCIX by +20.94. Overall, E2G-W had performed better

than Tapadóir and UCCIX by a significant margin of over +11, but had underperformed the other two models, IRIS and DGT-BPE16k, both by less than -7. Thus, E2G-W was slightly less competitive on the BLEU metric compared to IRIS and DGT-BPE16k.

On the other hand, our E2G-C model had outperformed every model on BLEU, except for DGT-BPE16k. E2G-C outperformed IRIS by +2.16, Tapadóir by +14.85, and UCCIX by +24.59. However, E2G-C had underperformed DGT-BPE16k by -2.57. Overall, E2G-C had the best results compared to other models, only falling short of DGT-BPE16k by -2.57 and had exceeded IRIS, UCCIX and Tapadóir.

We rank the performance of all examined models on BLEU in table 7.1. Based on the evaluated BLEU scores, we found that the best performing models, in order from best to worst, were: DGT-BPE16k, E2G-C, IRIS, E2G-W, Tapadóir, E2G-L, and then UCCIX. We found that our E2G-L model fails to contend with other models, but it remains competitive when translating text in the legal domain. Additionally, E2G-L had outperformed E2G-W on the Legal dataset, showing that the more in-depth legislative training data proved to increase the model's accuracy when translating in-domain text. E2G-W remained more consistent across datasets than E2G-L, but failed to achieve competitive results compared to E2G-C on any datasets. Furthermore, E2G-L had significantly underperformed on the Web dataset compared to all model scores, likely due to the low variety or count of commonly used words found in the Legal dataset. Overall, E2G-C was our best-performing and most consistent model on all of our datasets, showing that having a lot of unspecialized and specially curated training data can aid in a model's overall performance.

Model	BLEU Score
DGT-BPE16k	60.50
E2G-C	57.93
IRIS	55.77
E2G-W	54.28
Tapadóir	43.08
E2G-L	35.77
UCCIX	33.34

Table 7.1: Model performance ranking based on BLEU scores, ordered from best (top) to worst (bottom). Included E2G scores are the averaged scores from all our evaluated datasets.

7.2 TER

The TER metric measures the number of edits[1] a human translator would need to make to correct a machine-produced translation, divided by the average number of reference words. Hence, a lower TER score indicates a model produces translations closer to what a human translator would produce. Comparative models evaluated on the TER metric include: Tapadóir (Dowling, Cassidy, et al. 2015) and DGT-BPE16k (Lankford, Afli, and Way 2024).

We found E2G-L had significantly underperformed Tapadóir by +15.58 and DGT-BPE16k by +23.88. However, on the Legal dataset E2G-L had outperformed Tapadóir by -6.86 and only underperformed DGT-BPE16k by +6.44. For most cases E2G-L performs significantly worse compared to other models and is only moderately competitive when translating text in the legal domain. Our E2G-W model underperformed DGT-BPE16k by +6.16, but outperformed Tapadóir by -7.14. E2G-W outperformed Tapadóir almost as much as it underperformed compared to DGT-

[1]We discuss what an edit can refer to in chapter 5.3.2.

BPE16k. Our E2G-C model had significant outperformed Tapadóir by -10.24, while only underperformed DGT-BPE16k by +3.06. E2G-C is only slightly less competitive on TER compared to DGT-BPE16k but is more proficient than Tapadóir.

We rank the performance of all examined models on TER in table 7.2. Based on the evaluated TER scores, we found that the best performing models, in order from best to worst, were: DGT-BPE16k, E2G-C, E2G-W, Tapadóir, and then E2G-L. We found that E2G-L had significantly underperformed compared to other models, only achieving a competitive score on our Legal dataset. The performance of E2G-L was inconsistent overall, whereas E2G-W and E2G-C were much more consistent across all datasets. Furthermore, E2G-C had outperformed our other E2G models and only slightly underperformed DGT-BPE16k. Overall, our E2G-C model was our most consistent and competitive on TER, showing that including quality legislative data in training aids in generating translations closer to those of a human translator.

Model	TER Score
DGT-BPE16k	33.00
E2G-C	36.06
E2G-W	39.16
Tapadóir	46.30
E2G-L	56.88

Table 7.2: Model performance ranking based on TER scores, ordered from best (top) to worst (bottom). Included E2G scores are the averaged scores from all our evaluated datasets.

7.3 CHRF

CHRF calculates how well a predicted translation matches a reference translation, balancing the number of correct words with the number of missed words. A

higher CHRF score means a higher similarity between the prediction and reference. Comparative models evaluated on the CHRF metric include: IRIS (Arcan et al. 2016) and DGT-BPE16k (Lankford, Afli, and Way 2024).

Our E2G-L model had underperformed IRIS by -9.91 and DGT-BPE16k by -18.55. However, on our Legal dataset, E2G-L outperformed IRIS by +2.48, yet still underperformed DGT-BPE16k by -6.16. Thus, we found that E2G-L was much less competitive and significantly underperformed other models on CHRF, but remained competitive within the legal domain. However, E2G-W had outperformed IRIS by +3.37 but still underperformed DGT-BPE16k by -5.27. E2G-W is more competitive than IRIS but still fails to achieve competitive results compared to DGT-BPE16k. Our E2G-C model had outperformed IRIS by +5.36 and underperformed DGT-BPE16k by -3.28. Overall, we found that E2G-C had more competitive results to IRIS compared to E2G-W but similarly failed to outperform DGT-BPE16k.

We rank the performance of all examined models on CHRF in table 7.3. Based on the evaluated CHRF scores, we found that the best performing models, in order from best to worst, were: DGT-BPE16k, E2G-C, E2G-W, IRIS, and then E2G-L. We found that E2G-L fails to compete with other comparative models, only achieving similar results to other E2G models on our Legal dataset. E2G-W was our second-best model, outperforming IRIS and E2G-L, achieving relatively consistent results across all datasets. However, E2G-C was our best-performing model, having only underperformed DGT-BPE16k, and achieved consistent and competitive results across all our datasets. Overall, DGT-BPE16k and E2G-C are our best choice models for generating translations resembling human-produced translations, according to CHRF.

Model	CHRF Score
DGT-BPE16k	78.00
E2G-C	74.72
E2G-W	72.73
IRIS	69.36
E2G-L	59.45

Table 7.3: Model performance ranking based on CHRF scores, ordered from best (top) to worst (bottom). Included E2G scores are the averaged scores from all our evaluated datasets.

7.4 METEOR

The METEOR metric measures how well a predicted translation matches a reference by evaluating the correctness and order of words. A higher METEOR score represents that a model's predictions contained fewer missing words, more accurate word choices, and more precise ordering of words. The only comparative model evaluated on the METEOR metric was the IRIS model (Arcan et al. 2016).

All three of our E2G models outperformed IRIS by a significant margin. Our E2G-L model outperformed IRIS by +16.14, and significantly more by +33.5 on our Legal dataset, while E2G-W and E2G-C showed even larger gains of +33.05 and +36.1, respectively. Overall, there were improvements from all E2G models compared to IRIS, showing meaningful translation improvements in all our models.

We rank the performance of all examined models on METEOR in table 7.4. Based on the evaluated METEOR scores, we found that the best performing models, in order from best to worst, were: E2G-C, E2G-W, E2G-L, then IRIS. Overall, we found that all our E2G models outperform IRIS on METEOR by +16 or more, with E2G-C being the best model with a score of 76.83. Our findings suggest that our E2G models deliver more accurate translations in terms of word choice, sentence structure,

and overall fluency, aligning more closely with human translations compared to IRIS.

Model	METEOR Score
E2G-C	76.83
E2G-W	73.78
E2G-L	56.87
IRIS	40.73

Table 7.4: Model performance ranking based on METEOR scores, ordered from best (top) to worst (bottom). Included E2G scores are the averaged scores from all our evaluated datasets.

7.5 Key Findings

Overall, we found that our E2G-L model was inconsistent and underperformed all comparative models across all metrics and datasets, except on our Legal dataset where it was able to outperform E2G-W on all metrics. This is likely due to the specific legal training data, which may not have included many common everyday words or grammar. Our E2G-W model was our second best-performing model, scoring consistently across all metrics but not being able to out perform all comparative models. Furthermore, our E2G-C model was our best-performing and most consistent model across all metrics, scoring the best of our models across all metrics and was only found to underperform compared to DGT-BPE16k. The improvement in performance compared to our other models is likely due to the high-quality legal translations in combination with a lot of everyday vocabulary and grammar. E2G-C showed improvements on all metrics compared to previously created SMT models, showing that a modern NMT system can outperform previous SMT models when pre-trained and fine-tuned on English-Irish bilingual data.

The DGT-BPE16k model from Lankford, Afli, and Way 2024, was the only model to outperform our E2G-C model. This is likely due to their efforts in conducting hyperparameter optimization (HPO) on their models and incorporating BytePair-Encoding models with varying vocabulary sizes (from 4k-32k). HPO is the process of finding the best hyperparameters before training to maximize performance. Therefore, there is a possibility that if we conducted HPO on our models that they might be more competitive to, if not outperform, DGT-BPE16k.

The findings established in our research then allow us to respond to our initial research questions, discussed below.

RQ#1 How competitive is fine-tuning a pre-trained Transformer-based neural machine translation (NMT) model compared with other machine translation systems for translating between English and a low-resource language, such as Irish?

We found that our fine-tuned NMT models outperform previous SMT models on all metrics, and outperform most previous NMT models, only slightly underperforming, but remaining competitive, compared to the DGT-BPE16k model from Lankford, Afli, and Way 2024. Our E2G-L model is the only one found to be insufficiently competitive compared to other models overall, yet achieved competitive results on text in the legal domain. However, our E2G-W and E2G-C models remained competitive and demonstrated improvements across all metrics, despite a lack of architecture specialization or hyperparameter optimization.

RQ#2 How can the amount and quality of training data affect the performance of Transformer-based neural machine translation (NMT) models on BLEU, TER, CHRF, and METEOR evaluation scores for translating between English and a low-resource language, such as Irish?

We found that including more training data can improve results across all metrics, finding the best results from including both quality legal and web-crawled translations. E2G-L, trained solely on quality legal text, was the worst-performing model across all metrics, only outperforming E2G-W on our Legal dataset. We find that including a greater amount of "lower-quality" translation training data has a more significant effect on model performance compared to including a smaller amount of higher-quality training data. Thus, we find that including more training data, regardless of quality, helps improve results across all tested metrics. Additionally, further including quality translations showed improvements in the ability of our models to perform on domain specific datasets.

All code used for dataset preprocessing, model training, and evaluation is available in appendix B.2, ensuring full transparency and reproducibility of experiments. Furthermore, we publicly release our developed English-Irish bilingual datasets, Legal[2], Web[3], and Combined[4], along with our three fine-tuned models, E2G-L[5], E2G-W[6], and E2G-C[7], to the HuggingFace Hub. We release these resources to support further MT research for low-resource languages and promote future work for developing Irish-focused technologies.

[2] https://huggingface.co/datasets/jquigl/Processed_Legal_En-Ga
[3] https://huggingface.co/datasets/jquigl/Processed_Web_En-Ga
[4] https://huggingface.co/datasets/jquigl/Processed_Combined_En-Ga
[5] https://huggingface.co/jquigl/E2G-Legal
[6] https://huggingface.co/jquigl/E2G-Web
[7] https://huggingface.co/jquigl/E2G-Combined

8 Ethical Considerations

While developing our English to Irish MT models, we identified several ethical concerns. This chapter outlines these potential ethical issues identified through our research, from creating our datasets, model choice, and the processes for training and evaluating our models.

The first potential concern is the usage rights of the datasets we used to create our new datasets. We found that the Paracrawl and HPLT datasets were released under the Creative Commons CC0 license, and ELRC and Gaois were released under the Creative Commons Attribution 4.0 International license. However, the original DGT-TM and DCEP datasets don't state any copyright license, but both describe their usage conditions on their respective websites. However, the Kaggle version of the DGT-TM dataset was released using the Database Contents License (DbCL) v1.0.

Another potential concern, similar to one raised during our consultations with Irish-based organizations, is that our datasets weren't standardized to a single dialect of Irish. Thus, there's a possibility that generated translations are correct for one dialect of Irish but incorrect for the other two. It is unclear if our models have retained any dialect-specific focus or issues, and further work should be done to find any dialect flaws in our models, in addition to creating specialized models to be more equipped for translating into each dialect.

Our E2G-W and E2G-C models utilize web-crawled data, and consequently,

there's potential for our models to produce undesirable text when generating translations. Furthermore, because we fine-tune the M2M100 model, we do not know the exact data used for training, specifically Irish text. Therefore, there's potential for data overlap between the datasets used in their pre-training and the datasets we used for our fine-tuning.

Another concern for our results is that our comparisons between other models are not the most accurate, due to a mismatch of evaluation datasets and possible differences in the implementations of automatic metrics. To ensure the most accurate comparisons possible, it would be essential to have access to all models being compared and evaluate them using the same process on a new dataset that has not been seen by any model during training.

Finally, we did not conduct any human evaluations of the quality of the translations generated from our models. Thus, it's not completely apparent how well our models realistically translate from English to Irish. Further human evaluations are essential for fully assessing the quality of translations our models produce.

9 Conclusions

In this research, we attempted to answer whether modern Neural Machine Translation (NMT) systems are viable tools for translating low-resource languages, specifically for the Irish language. We created three English to Irish bilingual datasets, a Legal, Web, and Combined version, and used them for fine-tuning three separate English to Irish NMT transformer models, E2G-Legal (E2G-L), E2G-Web (E2G-W), and E2G-Combined (E2G-C). Foremost, we found that our E2G-L model was less competitive overall, but remained competitive within the legal domain, whereas in comparison, our E2G-W and E2G-C models were more competitive and demonstrated improvements across all evaluation metrics.

The results and findings from our research lay a foundation for future research for developing and improving Irish-based NMT systems and English-Irish bilingual datasets. Furthermore, we identified a high potential that conducting HPO techniques could improve the future quality of our models and their results. However, finding the most accurate comparisons between our models and others would require collecting all relevant models and conducting further automatic evaluations on a new dataset containing data not used during training any model, in addition to conducting further manual human evaluations. In conclusion, our models perform competitively compared to other models and demonstrate improvements across all evaluation metrics compared to the majority of previous English to Irish MT systems.

References

Arcan, Mihael et al. (May 2016). "IRIS: English-Irish Machine Translation System". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 566–572. URL: https://aclanthology.org/L16-1090 (cit. on pp. 7, 8, 27, 29, 31, 38, 39, 43, 44).

Banerjee, Satanjeev and Alon Lavie (June 2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72. URL: https://www.aclweb.org/anthology/W05-0909 (cit. on pp. 25, 30).

Bañón, Marta et al. (July 2020). "ParaCrawl: Web-Scale Acquisition of Parallel Corpora". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 4555–4567. DOI: 10.18653/v1/2020.acl-main.417. URL: https://aclanthology.org/2020.acl-main.417/ (cit. on p. 17).

Barry, James et al. (2022). *gaBERT – an Irish Language Model*. arXiv: 2107.12930 [cs.CL]. URL: https://arxiv.org/abs/2107.12930 (cit. on p. 9).

Bebis, G. and M. Georgiopoulos (1994). "Feed-forward neural networks". In: *IEEE Potentials* 13.4, pp. 27–31. DOI: 10.1109/45.329294 (cit. on p. 8).

Bradley, Michael (Jan. 2014). "Is it possible to revitalize a dying language? An examination of attempts to halt the decline of Irish". In: *Open Journal of Modern Linguistics* 04.04, pp. 537–543. DOI: 10.4236/ojml.2014.44047. URL: https://doi.org/10.4236/ojml.2014.44047 (cit. on pp. 2, 3).

Caballar, Rina Diane (2023). "How Indigenous Groups Are Leading the Way on Data Privacy". In: *Scientific American*. URL: https://www.scientificamerican.com/article/how-indigenous-groups-are-leading-the-way-on-data-privacy/ (cit. on p. 12).

Census 2022 (2022). *Census 2022: Fall in percentage of daily Irish speakers but greater proficiency among youth*. The Irish Times. URL: https://www.irishtimes.com/ireland/2023/05/30/census-2022-fall-in-percentage-of-daily-irish-speakers-but-greater-proficiency-among-youth/ (cit. on p. 3).

Chi, Zewen et al. (2024). "Can Pretrained English Language Models Benefit Non-English NLP Systems in Low-Resource Scenarios?" In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32, pp. 1061–1074. DOI: 10.1109/TASLP.2023.3267618 (cit. on p. 20).

Chiang, David (2005). "A hierarchical phrase-based model for statistical machine translation". In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL '05. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 263–270. DOI: 10.3115/1219840.1219873. URL: https://doi.org/10.3115/1219840.1219873 (cit. on p. 7).

Chung, Hyung Won et al. (2023). "UniMax: Fairer and More Effective Language Sampling for Large-Scale Multilingual Pretraining". In: *The Eleventh Interna-*

tional Conference on Learning Representations. URL: https://openreview.net/forum?id=kXwdL1cWOAi (cit. on pp. 21, 22).

Chung, Junyoung et al. (2015). *Gated Feedback Recurrent Neural Networks*. arXiv: 1502.02367 [cs.NE]. URL: https://arxiv.org/abs/1502.02367 (cit. on p. 8).

Conneau, Alexis et al. (July 2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747/ (cit. on p. 9).

Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media (2022). *Digital Plan for the Irish Language Speech and Language Technologies 2023-2027*. URL: https://assets.gov.ie/241755/e82c256a-6f47-4ddb-8ce6-ff81df208bb1.pdf (cit. on p. 13).

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/ (cit. on p. 9).

Doddington, George (Jan. 2002). "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics". In: pp. 138–145. DOI: 10.3115/1289189.1289273 (cit. on p. 31).

Dowling, Meghan, Lauren Cassidy, et al. (Nov. 2015). "Tapadóir: Developing a Statistical Machine Translation Engine and Associated Resources for Irish". In: (cit. on pp. 7, 27, 28, 38, 39, 41).

Dowling, Meghan, Teresa Lynn, et al. (Mar. 2018). "SMT versus NMT: Preliminary comparisons for Irish". In: *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*. Ed. by Chao-Hong Liu. Boston, MA: Association for Machine Translation in the Americas, pp. 12–20. URL: https://aclanthology.org/W18-2202 (cit. on p. 4).

Duh, Kevin et al. (May 2020). "Benchmarking Neural and Statistical Machine Translation on Low-Resource African Languages". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 2667–2675. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.325 (cit. on p. 5).

European Commission (Jan. 2022). en. URL: https://commission.europa.eu/news/irish-now-same-level-other-official-eu-languages-2022-01-03_en (cit. on p. 2).

European Commission, Directorate-General for Communications Networks, Content and Technology (2019). *English-Irish Website Parallel Corpus (Processed)*. Data set. URL: http://data.europa.eu/88u/dataset/elrc_2559 (cit. on p. 17).

Fan, Angela et al. (2020). *Beyond English-Centric Multilingual Machine Translation*. arXiv: 2010.11125 [cs.CL]. URL: https://arxiv.org/abs/2010.11125 (cit. on p. 21).

Gibert, Ona de et al. (2024). *A New Massive Multilingual Dataset for High-Performance Language Technologies*. arXiv: 2403.14009 [cs.CL]. URL: https://arxiv.org/abs/2403.14009 (cit. on p. 17).

Government of Ireland (2012). "Census of population 2011. high1: This is Ireland: highlights from census 2011, part 1". eng. In: Dublin: Stationery Office. ISBN: 978-1-4064-2650-2 (cit. on p. 3).

Hajlaoui, Najeh et al. (May 2014). "DCEP -Digital Corpus of the European Parliament". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari et al. Reykjavik, Iceland: European Language Resources Association (ELRA). URL: https://aclanthology.org/L14-1716/ (cit. on p. 16).

Hickey, Raymond (Mar. 2023). "Irish dialect classification". In: DOI: 10.34961/researchrepository-ul.24065574.v1. URL: https://researchrepository.ul.ie/articles/journal_contribution/Irish_dialect_classification/24065574 (cit. on p. 14).

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf. URL: https://doi.org/10.1162/neco.1997.9.8.1735 (cit. on p. 8).

Ji, Ziwei et al. (Mar. 2023). "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12, pp. 1–38. ISSN: 1557-7341. DOI: 10.1145/3571730. URL: http://dx.doi.org/10.1145/3571730 (cit. on p. 18).

Koehn, Philipp (2009). *Statistical Machine Translation*. Cambridge University Press (cit. on p. 7).

Koehn, Philipp, Hieu Hoang, et al. (June 2007). "Moses: Open Source Toolkit for Statistical Machine Translation". In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Ed. by Sophia Ananiadou. Prague, Czech Republic: Association for Computational Linguistics, pp. 177–180. URL: https://aclanthology.org/P07-2045 (cit. on p. 7).

Koehn, Philipp, Franz J. Och, and Daniel Marcu (2003). "Statistical Phrase-Based Translation". In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127–133. URL: https://aclanthology.org/N03-1017 (cit. on p. 7).

Lankford, Séamus, Haithem Afli, and Andy Way (Aug. 2021). "Machine Translation in the Covid domain: an English-Irish case study for LoResMT 2021". In: *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*. Ed. by John Ortega et al. Virtual: Association for Machine Translation in the Americas, pp. 144–150. URL: https://aclanthology.org/2021.mtsummit-loresmt.15/ (cit. on p. 9).

— (2024). *Transformers for Low-Resource Languages: Is Féidir Linn!* arXiv: 2403.01985 [cs.CL]. URL: https://arxiv.org/abs/2403.01985 (cit. on pp. 5, 10, 27–29, 38, 39, 41, 43, 46).

Li, Lan, C. Brugha, and Mary Gallagher (Dec. 2017). "Protecting Endangered Languages: The Case of Irish". In: *Studies in Arts and Humanities* 3, pp. 109–130. DOI: 10.18193/sah.v3i2.110 (cit. on p. 6).

Liu, Vivian and Yiqiao Yin (2024). *Green AI: Exploring Carbon Footprints, Mitigation Strategies, and Trade Offs in Large Language Model Training*. arXiv:

2404.01157 [cs.CL]. URL: https://arxiv.org/abs/2404.01157 (cit. on pp. 22, 24).

Magueresse, Alexandre, Vincent Carles, and Evan Heetderks (2020). *Low-resource Languages: A Review of Past Work and Future Challenges*. arXiv: 2006.07264 [cs.CL]. URL: https://arxiv.org/abs/2006.07264 (cit. on pp. 1, 20).

Ní Chiaráin, Neasa et al. (2023). "Filling the SLaTE: examining the contribution LLMs can make to Irish iCALL content generation". In: *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pp. 176–181. DOI: 10.21437/SLaTE.2023-34 (cit. on p. 4).

NLLB Team et al. (2022). *No Language Left Behind: Scaling Human-Centered Machine Translation*. arXiv: 2207.04672 [cs.CL]. URL: https://arxiv.org/abs/2207.04672 (cit. on p. 21).

Ogueji, Kelechi, Yuxin Zhu, and Jimmy Lin (2021). "Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages". en. In: *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 116–126. DOI: 10.18653/v1/2021.mrl-1.11. URL: https://aclanthology.org/2021.mrl-1.11 (cit. on pp. 22, 23).

Papineni, Kishore et al. (2002). "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://doi.org/10.3115/1073083.1073135 (cit. on pp. 25, 26).

Popović, Maja (Sept. 2015). "chrF: character n-gram F-score for automatic MT evaluation". In: *Proceedings of the Tenth Workshop on Statistical Machine Transla-*

tion. Ed. by Ondřej Bojar et al. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395. DOI: 10.18653/v1/W15-3049. URL: https://aclanthology.org/W15-3049/ (cit. on pp. 25, 28, 29).

Popović, Maja (Sept. 2017). "chrF++: words helping character n-grams". In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 612–618. DOI: 10.18653/v1/W17-4770. URL: https://aclanthology.org/W17-4770 (cit. on p. 28).

Post, Matt (Oct. 2018). "A Call for Clarity in Reporting BLEU Scores". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, pp. 186–191. URL: https://www.aclweb.org/anthology/W18-6319 (cit. on pp. 27, 28).

Pyysalo, Sampo et al. (2020). *WikiBERT models: deep transfer learning for many languages*. arXiv: 2006.01538 [cs.CL]. URL: https://arxiv.org/abs/2006.01538 (cit. on p. 9).

Raffel, Colin et al. (2023). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. arXiv: 1910.10683 [cs.LG]. URL: https://arxiv.org/abs/1910.10683 (cit. on pp. 21, 22).

Snover, Matthew et al. (Aug. 2006). "A Study of Translation Edit Rate with Targeted Human Annotation". In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, pp. 223–231. URL: https://aclanthology.org/2006.amta-papers.25/ (cit. on pp. 25, 27).

Steinberger, Ralf et al. (2013). *DGT-TM: A freely Available Translation Memory in 22 Languages*. arXiv: 1309.5226 [cs.CL]. URL: https://arxiv.org/abs/1309.5226 (cit. on p. 16).

Tran, Khanh-Tung, Barry O'Sullivan, and Hoang D. Nguyen (2024). *UCCIX: Irish-eXcellence Large Language Model*. arXiv: 2405.13010 [cs.CL]. URL: https://arxiv.org/abs/2405.13010 (cit. on pp. 9, 10, 27, 38, 39).

Vallor, Shannon (June 2024). *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. en. 1st ed. Oxford University PressNew York. ISBN: 978-0-19-775906-6. DOI: 10.1093/oso/9780197759066.001.0001. URL: https://academic.oup.com/book/56292 (cit. on p. 17).

Vaswani, Ashish et al. (2017). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762 (cit. on pp. 9, 10, 20).

Xue, Linting et al. (2021). *mT5: A massively multilingual pre-trained text-to-text transformer*. arXiv: 2010.11934 [cs.CL]. URL: https://arxiv.org/abs/2010.11934 (cit. on pp. 21, 22).

A Organizations

A.1 List

This section describes the many Irish organizations that were contacted to collect stakeholder feedback on the usage of LLMs and how they could be best utilized to aid in the preservation, promotion, and/or education of the Irish language. Table A.1 lists the organizations contacted, the date of contact, if a response was received, and the outcome of their response, if applicable.

Organization Name	Contact Details	Date of Contact	Response Received	Outcome
Conradh na Gaeilge	Website: cnag.ie/en/ Email: eolas@cnag.ie	Apr 12, 2024	N	N/A
Gaeloideachas	Website: gaeloideachas.ie Email: eolas@gaeloideachas.ie	Apr 16, 2024	Y	Conducted Meeting
Gael Linn	Website: gael-linn.ie/en/ Email: eolas@gael-linn.ie	May 15, 2024	N	N/A
An Chomhairle um Oideachas Gaeltachta & Gaelscolaíochta	Website: www.cogg.ie/en/ Email: eolas@cogg.ie	May 21, 2024	Y	Advised to contact Údarás na Gaeltachta & Roinn na Gaeltachta
Údarás na Gaeltachta	Website: udaras.ie/en/ Email: eolas@udaras.ie	May 28, 2024	N	N/A
Roinn na Gaeltachta (TCAGSM) - Community and Language Support Programme	Website: gov.ie/en/organisation/ Email: ctpt@tcagsm.gov.ie	May 28, 2024	Y	Conducted Meeting
Royal Irish Academy	Website: www.ria.ie Email: info@ria.ie	May 30, 2024	N	N/A
Foras na Gaeilge	Website: forasnagaeilge.ie/ Email: eolas@forasnagaeilge.ie	Jun 3, 2024	N	N/A
Comhar Naíonraí na Gaeltachta	Website: comharnaionrai.ie/ Email: eolas@cnng.ie	Jun 3, 2024	N	N/A
Gaelscoileanna Teo	Website: gaelscoileanna.ie/en/ Email: oifig@gaelscoileanna.ie	Jun 5, 2024	N	N/A

Table A.1: Irish Organizations that were contacted for consultations regarding the use of AI/LLMs for the Irish Language

A.2 Questions

This section outlines the standard questions asked during our meetings with Irish organizations, but does not cover everything discussed. The questions were asked sequentially, and their responses were recorded after obtaining consent from the interviewee(s). However, we do not disclose any specifics discussed during our conversations, and all significant aspects from these meetings were discussed and summarized in chapter 7.

1. Do you have any experience using AI tools in Irish? If yes, why and how well did it work?

2. Are there any common challenges people face when communicating through text in Irish?

3. Can you describe any specific nuances or subtleties in Gaeilge that might make it difficult for non-native speakers or simple algorithms to understand?

4. How important is context in understanding the meaning of words or phrases in Irish?

5. How much should Scottish-Gaelic be considered when developing our AI tool?

6. What are your opinions about the preservation of cultural/linguistic diversity in the digital age, especially in the context of AI / AI models?

7. Would you prefer an AI/language model that prioritizes accuracy or one that prioritizes generating more natural-sounding text? (If natural; then won't be described as being correct)

8. Are there any privacy concerns –that you can think of– regarding the use of Irish language data to train language models?

9. How do you think language models could be used to benefit society in your language community?

10. What are your thoughts on the ethical implications of developing AI models for Irish, especially considering issues like bias and misinformation?

11. Are there any types of content or topics that tend to be produced or discussed more often than others in Irish?

12. Is there anything I didn't ask about that I should have or that you think might be relevant or important?

13. Are you aware of anyone else who might be interested in talking to me that you are aware of?

B Code

B.1 Examples

```python
# Method to drop rows given a condition function
def drop_rows_custom(df, condition_func):
    mask = df.apply(condition_func, axis=1)
    return df[~mask]

# Method to drop rows where the English column contains profanity
def condition(row):
    return profanity.contains_profanity(row['en'])

# Print original dataset length
print(len(df))
# Use the custom method
df_filtered = drop_rows_custom(df, condition)
# Print new length
print(len(df_filtered))
```

Listing B.1: Python example of removing profane entries using better-profanity.

```python
df.dropna(axis=0, how='any')
df.drop_duplicates(subset=['en'], keep='first', inplace=False)
```

Listing B.2: Python example of removing NULL and duplicate entries.

```python
max = len(df)
for i in range(0, max):
  item = df.iloc[i]
  try:
    # English regex
    # Removing leading newline characters:
    en = re.sub(r"\n", "", item['en'])
    # Removing leading digit(s)/special character(s):
    en = re.sub(r"^(?: ?\d+?(\d[a-zA-Z]+\D ?)|\w?\d*\W+)+", "", en)
    # Replacing English entry with our processed version
    df.loc[i, 'en'] = en

```

```
    # Irish regex
    # Removing leading newline characters:
    ga = re.sub(r"\n", "", item['ga'])
    # Removing leading digit(s)/special character(s):
    ga = re.sub(r"^(?: ?\d+?(\d[a-zA-Z]+\D ?)|\w?\d*\W+)+", "", ga)
    # Replacing Irish entry with our processed version
    df.loc[i, 'ga'] = ga
  except:
    print(item)
```

Listing B.3: Python example of our process for preprocessing entries using regular expressions, removing leading digits and special characters.

Input	**Result**
“1° Lorem ipsum dolor sit amet”	“Lorem ipsum dolor sit amet”
“\nLorem ipsum dolor sit amet”	“Lorem ipsum dolor sit amet”
“\n1 Lorem ipsum dolor sit amet”	“Lorem ipsum dolor sit amet”
“1. Lorem ipsum dolor sit amet”	“Lorem ipsum dolor sit amet”
“1) Lorem ipsum dolor sit amet”	“Lorem ipsum dolor sit amet”
“\n 1° Lorem ipsum dolor sit amet”	“Lorem ipsum dolor sit amet”
“ Lorem ipsum dolor sit amet”	“Lorem ipsum dolor sit amet”
“7th parliamentary term”	“7th parliamentary term”
“7ú téarma parlaiminteach”	“7ú téarma parlaiminteach”
“1 1° Lorem ipsum dolor sit amet 1 ”	“Lorem ipsum dolor sit amet 1 ”
“(Lorem ipsum)”	“(Lorem ipsum)”
“(Lorem ipsum)”	“Lorem ipsum)”

Table B.1: Examples of how our dataset preprocessing can affect translation entries.

B.2 Resources

Example code used for our dataset preprocessing, specifically for the DGT-TM dataset: https://tinyurl.com/E2Gpreprocess.

Example code used for training our E2G-Legal model can be seen here: https://tinyurl.com/E2G-train. Note that the same process was used for our other models, just changing the dataset used for training.

Example code used for evaluating our E2G-Legal model can be seen here: https://tinyurl.com/E2G-eval. Note that the same process was used for our other models, just changing the model loaded for evaluation.