Final Year Project: Facilitating Access to English-Irish COVID-19 Information via Neural Machine Translation

Jack Boylan (17774291)

Dublin City University jack.boylan25@mail.dcu.ie

Abstract. Interlingual communication has always been important, but if there was ever any doubt about this, the ongoing COVID-19 pandemic has shown it to be particularly true. New information is being released every day about travel restrictions, economic impacts, public health and safety procedures, and potential vaccines. Barriers of language should be removed where possible to ensure that everyone can stay up-to-date and safe, regardless of how many speak the language. Previous work has been carried out to aid rapid translation of COVID-19 related material in several major languages. However, absent among the languages served is Irish.

The case of providing effective Machine Translation (MT) for Irish is a particularly important problem, not only in light of the global pandemic; Irish faces the expiration of a derogation at the end of 2021, at which point it will become a full working language of the European Union, and thus will be required to translate all legislation enacted from that point onwards. There is an urgent need for solutions to be created in this area to aid in the enormous translation workload that will soon meet Ireland.

In this project, we have created a Neural Machine Translation (NMT) system for English into Irish using freely available data. This system can be accessed worldwide via a web interface, extending existing work in making COVID-19 related materials accessible in other languages via online applications. We have also compared our model to the leading MT systems available online today.

Keywords: Neural Machine Translation · COVID-19 · Machine Learning

1 Motivation and Background

For my Third Year INTRA internship, I spent 7 months working at Iconic Translation Machines.[1] Iconic is a language technology software company that supplies bespoke MT systems to a global client base for a variety of specific use-cases. During this time, I was exposed to the power of machine learning to break down

[1] https://iconictranslation.com/

barriers between languages. One standout example of this was the PRINCIPLE Project, a European Commission-funded project aimed at providing language data to improve translation quality in the European Digital Service Infrastructures (DSIs) of eJustice and eProcurement via domain-specific NMT.[2] Iconic are focusing on developing MT engines in these domains for low-resource languages Croatian, Icelandic, Irish, and Norwegian. I gained a particular interest in Irish MT and the implementation of state-of-the-art technologies in approaching the challenge. In my spare time, I tried building my own Transformer model [27] for English into Irish using Tensorflow.[3] This allowed me to become familiar with the necessary steps of MT, such as preprocessing data, tokenising text, configuring model parameters, and evaluating translation quality. For my final year project, I wanted to further investigate the tools and technqiues used in state-of-the-art MT systems.

Over the course of the COVID-19 pandemic, we have seen countries struggle in the face of heavy restrictions and increasingly strict safety protocols. Civil unrest in response to these measures has been manifested in protests across the country and the illegal operation of services for businesses that are temporarily required to close. Some countries[4][5] have handled the situation remarkably well, due in part to excellent public communication of up-to-date information, facilitating a greater understanding amongst the public about what measures have been put in place and how such measures are protecting public health.

In some scenarios where language has become a barrier to effective communication, rapidly developed MT systems have provided an excellent gateway for accessing the latest COVID-19 material [28].

In the cited paper, MT systems were built for high-resource languages English, French, Italian, German and Spanish. These solutions have the ability to reach an enormous number of people who can easily access information between any of the aforementioned languages. However, communities of minority languages do not have this opportunity, and so they are at risk of falling behind on the latest protocols or important announcements. One such example closer to home is Irish.

Providing up-to-date translations of COVID-19 related material in Irish presents a significant challenge. Unfortunately, due to the country's low population and the scarcity of qualified Irish translators, meeting this challenge within a short time frame will prove very difficult. Additional translators are being recruited, and new translation courses are being introduced throughout the country's colleges to meet future demand; however, these solutions will take considerable time and funds to complete.

Existing research has shown that MT for English-Irish is a challenging problem, due to low resources and complex language structure [6]. Statistical Ma-

[2] https://bit.ly/3bAgPgP

[3] https://colab.research.google.com/drive/1efwxpdfoXQiMKgHDOCqXDXRfOr-xOXMY?usp=sharing

[4] https://time.com/5851633/best-global-responses-covid-19/

[5] https://www.movehub.com/blog/best-and-worst-covid-responses/

chine Translation (SMT) approaches have shown reasonable performance in the past [9] [3], often outperforming NMT iterations [8], although the fluency aspect of output using these systems has struggled. These initial studies investigating NMT for English-Irish have not proven to be as promising as one might hope. Such results may possibly be due to the lack of large bodies of bilingual corpora and the challenging structure of the Irish language. For example, more recent research has opened a more positive outlook on NMT for Irish, primarily through the addition of backtranslation techniques [5].

The latest evolution of NMT is the Transformer architecture [27], which currently holds top position in encoder-decoder sequence-to-sequence networks. This model makes use of multi-head attention mechanisms within an encoder and decoder stack, and removes the need for recurrence and convolutions. A graphical representation of the architecture is given in Figure 1. Since its creation, it has quickly become the model of choice for NMT, repeatedly outperforming previous architectures.

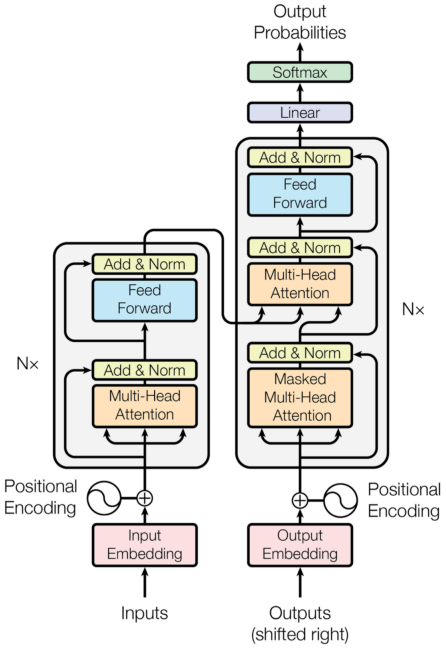

Fig. 1: Transformer Model Architecture [27].

2 Collection and Preparation of Datasets

Our objective in this project is to develop a NMT system for the English-Irish language pair that is capable of effectively translating COVID-19 related material. We accomplish this through use of English-Irish parallel corpora that are

freely available online. Additional synthetic data is generated through monolingual corpora to provide more input for model training.

The use of NMT over SMT means that much more training data is required. NMT models are also more sensitive than SMT models to noise during training and so our larger dataset must also be in a cleaner format [17]. Irish is considered a low-resource language [2], which means that in order to gather a sufficient amount of input data, we make use of data augmentation techniques such as backtranslation [24], an approach whereby monolingual corpora in the target language are translated back into source languages through an MT system, and this output is used to help train another MT system. This technique will be discussed further in Section 2.1. The results from these techniques are promising, although the ratio of authentic to synthetic data must be carefully monitored in order to achieve optimal performance [21] [7]. The approach has also appeared to be more effective in the GA-EN direction, while EN-GA has seen less of an improvement, particularly due to the task of translating into a more morphologically rich language with greater possible inflected forms (only one of which will be used as the correct reference translation), as well as a lack of monolingual corpora and domain-specific data. Iterative backtranslation [12] is also examined, as it has also been shown to lend a performance boost in other language pairs, but has yet to be implemented in an English-Irish MT system according to my prior research.

Due to the small population of Irish speakers, the data available for English-Irish translation is relatively low when compared to languages such as French, German or Spanish. Datasets available for public use are often either small or suffer from issues of poor translation and/or misalignment. More scarce still is the availability of English-Irish data in the domain of COVID-19 due to its rapid emergence.

Data sources used throughout this project can be found in Table 1.

Table 1: Breakdown of data sources.

Source Name	Data Type	Num. sentences
ParaCrawl 7	Parallel	2,671,527
Citizens Information Bilingual Web-Corpus	Parallel	10,297
OpenSubtitles (OPUS)	Monolingual English	486,609
Legal Acts of Ireland	Monolingual Irish	96,225
Tatoeba Project	Parallel	698
Irish Wikipedia	Monolingual Irish	249,876
Web Scraping	Monolingual Irish	18,418
TAUS	Monolingual English	879,926
EMEA	Monolingual English	1,108,752

The ParaCrawl dataset [4] is a set of large parallel corpora to/from English for all official EU languages created through a broad web crawling effort. Through

each iteration of ParaCrawl, more data is accumulated for each language pair, and more advanced tools are implemented to clean text and improve overall data quality. In our experiments we are making use of the ParaCrawl 7 English-Irish corpus.

ELRC-share[6] is a repository for language resources and tools used in translation. This repository was used to gather parallel sentences from the Citizens Information Bilingual Web-Corpus, Irish Monolingual Corpus from the contents of the Department of Health website,[7] COVID-19 Social Questions - Health corpus, the Conference of the International Association of Language Commissioners (Comhdháil Cumann Idirnáisiúnta na gCoimisineirí Teanga) and the Tatoeba parallel dataset.

2.1 Using Backtranslated/Iteratively backtranslated data

Monolingual corpora in Irish will be used for backtranslation (Figure 2). Backtranslation, proposed by Sennrich et al., 2015 [24], makes use of a target monolingual corpus – in this case Irish – and translates the text back into the source language – English – via another MT system. In our case, GA-EN translation is an easier direction, as we move from a morphologically rich language to a comparatively poor language. Systems such as Google Translate provide reasonable performance, and may be used for such purposes.

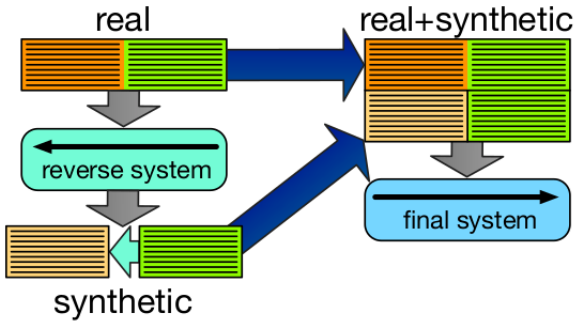

Fig. 2: Backtranslation overview [12].

The Open Parallel Corpus (OPUS)[8] data was acquired in monolingual English extracted from an English-Turkish dataset and translated to Irish using an existing MT engine to provide a source of synthetic data for further experimentation. Irish monolingual data or data with no/poor quality English translation was passed through a backtranslation script in an effort to gain additional data.

[6] https://elrc-share.eu/
[7] https://www.gov.ie/en/organisation/department-of-health/
[8] http://opus.nlpl.eu/

Additional monolingual data was acquired from Irish Wikipedia,[9] which can also be used as an augmented data source.

Web scraping of public health websites was performed to retrieve monolingual Irish data that could serve as a synthetic source of up-to-date COVID-19 related parallel sentences. This collection was carried out using the ScraPy[10] library in accordance with ethical web scraping protocols.

The TAUS Corona[11] dataset is a large collection of corpora containing parallel data in the COVID-19 domain. While no such dataset is available for English-Irish, we can use the English monolingual data available in the English-Spanish corpus to provide synthetic data. This dataset consists of 879,926 sentences, before preprocessing. A similar resource was gathered from the EMEA[12] Corpus, which provides data in the medical domain for several major languages. We can employ methods of iterative backtranslation demonstrated by [12] (Figure 3) to take the English side of these corpora, and create translations in Irish. Using these target sentences, we can perform backtranslation as previously described to gain some additional domain-specific examples. It was noted that these sentences were not of excellent quality, as they have been passed through several layers of MT. However, it was of interest to observe if they provided any benefit to our MT system performance, as has been demonstrated for other language pairs.

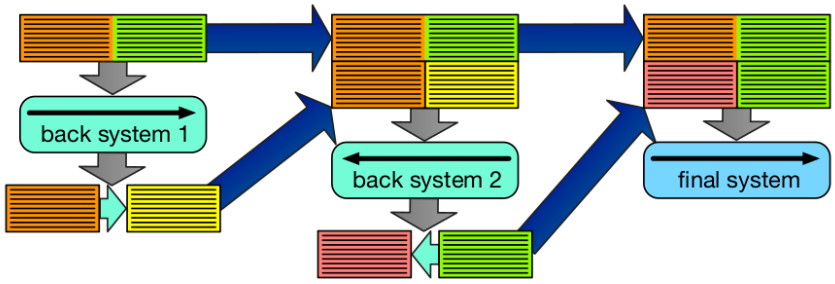

Fig. 3: Iterative backtranslation overview, [12].

2.2 Data Preparation

While a reasonable body of text is available from the above mentioned resources, the number of examples is reduced significantly after preprocessing sentences for noise and proper alignment. Preprocessing involves removing sentences less than 15 or more than 300 characters in length and cleaning sentences of special

[9] https://ga.wikipedia.org/wiki/Vicip\%C3\%A9id
[10] https://scrapy.org/
[11] https://md.taus.net/corona
[12] http://opus.nlpl.eu/EMEA.php

characters and hyperlinks, among other processes suggested to create a robust NMT model [11].

A common issue encountered throughout the larger datasets was finding the English source sentence copied as the Irish translation. In an attempt to alleviate this problem, the langdetect[13] package was used to verify that parallel sentences actually consisted of an English source sentence and Irish target translation.

None of the data is gathered from personal or sensitive sources, and so there are no ethical risks to be considered. Experiments were carried out by training models using varying ratios of authentic and synthetic data to identify optimal performance on both in- and out-of-domain data.

Byte-pair encoding [25] was carried out on the data before training the model to reduce memory consumption and improve training and inference speed. This process was achieved through use of the Sentencepiece [18] tool developed by Google. BPE allows us to break down words into common subword tokens, which may then be used to construct longer and more uncommon words.

3 Experiments and Results

3.1 Model Set Up

The Transformer model was built using the OpenNMT-tf [16] framework. This is a Python-based toolkit built upon the Tensorflow platform, used to create state-of-the-art MT models [1]. In all experiments, the same model is used as described by Vaswani et al. [27]. This means there are 6 layers of encoders and decoders, with 8 heads of attention.The weight matrix is shared amongst the embedding layers and the default inter-layer dropout rate is 0.1. Model dimension is set to 512, data is read in batches of 32 and the model is given 8,000 warm up steps at the beginning of the training stage. An Adam optimiser [14] is used during training to provide a method of updating parameters through first-order gradient descent. The training is evaluated at regular intervals on a validation set of 5,000 sentences. At each interval the BLEU [20] score will be obtained by computing n-gram overlap between the reference and hypothesis.

The BLEU score recorded for these intervals must improve by at least 0.5 every 2 evaluations or the model will finish early. If the early stopping criterion is not met, the model will continue to a maximum step value of 1,000,000. A number of recent model checkpoints will be saved during each experiment, and after training the parameters from the last 5 checkpoints will be averaged to create the final model, as this has often been shown to improve model robustness and avoid overfitting [19].

3.2 Experiment Results

Experiments were conducted using various ratios of authentic, backtranslated and iteratively-backtranslated corpora in order to find an optimal balance be-

[13] `https://pypi.org/project/langdetect/`

tween authentic and synthetic data. The final model was saved to cTranslate2[14] format, which is more suitable than Tensorflow's SavedModel format for online deployment due to its lighter weight, faster inference speeds and fewer dependencies [15].

Each model will be tested on an individual unseen test set of 20,000 EN-GA sentences. Another test set of 1,500 sentences containing COVID-19 related material is also used for testing at this time. Previous research [26] has suggested that metrics such as BLEU that rely on word-based computations are not suitable measures of performance when comparing MT systems, so we will also be using chrF [22], a character-based metric which compares up to 6 character n-grams. BLEU and chrF scores will be recorded during testing and this will be used to assess model performance. The test results are reported in Table 2.

Table 2: Breakdown of test scores for authentic, backtranslated* (BT), and iteratively backtranslated** (ITB) data.

Title	Test Data		COVID-19 Test Data	
	BLEU	chrF	BLEU	chrF
Parallel Only	**56.5**	**74.0**	28.4	51.75
Parallel +50k (BT)*	56.0	73.78	28.9	52.22
Parallel +100k (BT)	54.6	73.15	28.9	52.08
Parallel +250k (BT)	54.8	73.28	29.0	52.14
Parallel +500k (BT)	54.5	73.11	29.0	52.91
Parallel +1M (BT)	54.5	73.18	**29.5**	**52.99**
Parallel +ALL(BT)	53.1	72.29	28.7	52.55
*Parallel +50k (IBT**)*	54.8	73.22	29.0	52.27
Parallel +100k (IBT)	55.1	73.36	28.6	52.14
Parallel +250k (IBT)	54.9	73.23	28.9	52.06
Parallel +500k (IBT)	53.1	72.34	28.9	52.70
Parallel +1M (IBT)	53.7	72.71	29.4	52.67
Parallel +ALL(IBT)	54.0	72.78	28.1	52.25

We note good performance on the test sets across all models, with a noticeable drop in quality for the COVID-19 domain test set. However, these scores indicate that the models may still perform reasonably well for our use case. Our baseline Parallel Only model has achieved the best scores for our larger general test set, but has been beaten by the Parrallel + 1M (BT) model on the in-domain COVID-19 test set. We will choose the Parallel + 1 Million backtranslated sentences model as our model to investigate further, due to the improved BLEU and chrF scores on our COVID-19 test set. Using the Wilcoxon rank sum test [29] to compare our models, we find that the results for the COVID-19 test set from our Parrallel + 1M (BT) model is statistically significant when compared

[14] `https://github.com/OpenNMT/CTranslate2`

to the baseline Parallel Only model for $p - value < 0.05$. Significance testing was carried out using the testing scripts made available following the paper by Dror et al., 2018 [10]

3.3 How does this engine compare to current online MT systems?

We have compared our models against one another, but we are also interested in how they may perform in comparison to current online MT services. We will test Google Translate and Microsoft Translator on our larger test set and COVID-19 test set. In Table 3, we report the results of our tests.

Table 3: Comparing our best model against Leading Online Translators.

Title	Test Data		COVID-19 Test Data	
	BLEU	chrF	BLEU	chrF
Parallel +1M (BT)	**54.5**	**73.18**	**29.5**	**52.99**
Microsoft Translator	33.6	60.52	21.5	49.18
Google Translate	30.1	58.87	19.3	46.40

We can see strong performance by our engine for the larger test set in terms of BLEU scores relative to the Google Translate and Microsoft Translator; the model trained on parallel data and 1 million backtranslated sentences achieves a 62 per cent greater score for this metric than the next best Microsoft Translator. However, examining chrF scores tells us that this gap in performance may not be as significant as the BLEU scores would indicate; a 20 per cent higher value is obtained in this case. A significant drop in performance is also noted across the board when testing with the COVID-19 text data. In terms of BLEU score, Microsoft Translator is better than Google Translate, achieving 3.5 more points on the larger test set and 2.2 points on the COVID-19 test set. Meanwhile, our best model obtains 8 points higher BLEU score in this case. This may be due in part to the fact that the BLEU metric was our early stopping criterion during training, whereas the training process for the other engines is unknown.

Similar to the larger test set, a closer competition emerges when comparing chrF scores. Our best model performs 12.6 points higher on the larger test set than the next best Microsoft Translator, which is 1.7 points higher than Google Translate. The COVID-19 test set chrF scores are 52.99, 49.18 and 46.40 for our best model, Microsoft Translator and Google Translate, respectively. These scores may indicate the difficulty of translating this in-domain material.

Making use of the Wilcoxon test again, we can confirm that the results from our models are statistically significant for $p - value < 0.05$ when compared to the next best Microsoft Translator on both the larger test set and the COVID-19 test set.

3.4 Human Evaluation

Our models exhibit good performance under automatic evaluation metrics, but we primarily aim to achieve good performance in real-world use. Below we will compare several different translation samples of sentences from the COVID-19 test set to experience end-user output and identify some issues the model may present. We will again compare Google Translate and Microsoft Translator to our engine.

All models appear to perform reasonably well for these examples. A common observation throughout the translations is that each system often focuses on fluency over total adequacy of output given the source sentence. This is one of the challenges faced by NMT systems noted by [17]. This means that, while not completely incorrect, the output of these systems often diverges from the reference in terms of word order (example (1)) and verbs used (example (3)). However, the overall meaning is often kept intact.

The translation of English-Irish is difficult due to the increased morphological complexity of the target language compared to the source. This can be seen in the examples where the incorrect tense is used or grammatical practices present only in Irish are not employed by the MT system.

In example (2), our model translates the source sentence, leaving out the starting "As hospital prescribers". This is an important element of the sentence and should be preserved as seen in the other systems' outputs.

In example (5), our model translates the word 'immunised' to the Irish for "immunisation". The other systems translate the word correctly. Our model and Google Translate start the same example with the phrase "Ní amháin go gcosnaíonn" ("Not only does it protect") instead of "Cosnaíonn díolúine trí vacsaíniú", an example of where the fluency of the models outweighs the focus for direct translation.

In cases where rare words are encountered (examples (1) and (7)), the word is occasionally left untranslated by one or more systems. This is normally the issue with chemical compounds or equations that the system has never seen before. However, we also see this happen in example (3) with the word "Infographic" left untranslated by our model and Google Translate.

Table 4: Outputs of MT engines for various inputs (1) - (4).

Origin	Sentence
(1) Source	The bacterium that was isolated from Mohammed 's bloodstream was resistant to many antibiotics including the last source antibiotics a class of last line antibiotics called carbapenems .
Reference	Bhí an baictéar a leithlisíodh as sruth fola Mohammed frithsheasmhach in aghaidh a lán antaibheathach lenar áiríodh antaibheathaigh na rogha deireanaí is iad sin aicme antaibheathach ar a dtugtar carbaipeineimí .
Our Model	Bhí an baictéar a bhí scoite amach ó shruth fola Mohammed frithsheasmhach in aghaidh antaibheathach go leor lena n-áirítear antaibheathaigh foinse deiridh aicme antaibheathach líne dheiridh ar a dtugtar carbapenems.
Microsoft	Bhí an baictéar a bhí scoite amach ó sruth fola Mohammed frithsheasmhach in aghaidh go leor antaibheathach lena n-áirítear antaibheathaigh na foinse deireannaí aicme antaibheathach líne deireanach ar a dtugtar carbapenems .
Google	BHI ar baictéar a Glossary scoite gcéill Ó sruth fola Mohammed frithsheasmhach in Aghaidh dul Leor antaibheathach lena n-áirítear antaibheathaigh na Foinse deireannaí Aicme antaibheathach líne Deireanach Cruinniú a dtugtar carbapenems
(2) Source	As hospital prescribers your tasks related to improving antibiotic use include
Reference	Áirítear iad seo a leanas leis na cúraimí atá ort mar oideasóir ospidéil maidir le húsáid antaibheathach a fheabhsú
Our Model	Áirítear ar do chuid cúraimí a bhaineann le húsáid antaibheathach a fheabhsú
Microsoft	Mar oideasóirí ospidéil tá do chúraimí a bhaineann le húsáid antaibheathach a fheabhsú
Google	Mar oideasóirí ospidéil tá do chuid cúraimí a bhaineann le húsáid antaibheathach a fheabhsú
(3) Source	Infographics explaining the process of antibiotic resistance the relation between antibiotic consumption and antimicrobial resistance the recent trend of growing resistance to last line antibiotics .
Reference	Bileog eolais don phobal i gcoitinne agus do shaineolaithe maidir le frithsheasmhacht in aghaidh antaibheathach sa chúram príomhúil agus in ospidéil
Our Model	Infographics ag míniú an phróisis frithsheasmhachta in aghaidh antaibheathach an gaol idir tomhaltas antaibheathach agus frithsheasmhacht in aghaidh ábhar frithmhiocróbach an treocht le déanaí maidir le frithsheasmhacht in aghaidh antaibheathach líne dheiridh a mhéadú .
Microsoft	Grafaicí faisnéise a mhíníonn próiseas na frithsheasmhachta in aghaidh antaibheathach an gaol idir tomhaltas antaibheathach agus frithsheasmhacht in aghaidh ábhair fhrithmhiocróbacha an treocht le déanaí maidir le frithsheasmhacht in aghaidh antaibheathach líne deiridh a fhás.
Google	Infographics ag míniú an phróisis frithsheasmhachta in aghaidh antaibheathach an gaol idir tomhaltas antaibheathach agus friotaíocht frithmhiocróbach an treocht le déanaí maidir le friotaíocht a mhéadú in aghaidh antaibheathach líne dheiridh.
(4) Source	Avoid unnecessary antibiotic prophylaxis .
Reference	Próifiolacsas neamhriachtanach antaibheathach a sheachaint .
Our Model	Seachain próifiolacsas antaibheathach gan ghá .
Microsoft	Seachain próifiolacsas antaibheathach neamhriachtanach .
Google	Seachain próifiolacsas antaibheathach gan ghá.

Table 5: Outputs of MT engines for various inputs (5) - (7).

Origin	Sentence
(5) Source	Immunity through vaccination protects not only the immunised individual but also protects unvaccinated people in the community such as infants who are too young to be vaccinated .
Reference	Cosnaíonn díolúine trí vacsaíniú ní hamháin an duine aonair imdhíonta ach cosnaíonn sé daoine neamh-vacsaínithe sa phobal ar nós naíonáin atá ró-óg le vacsaíniú.
Our Model	Ní amháin go gcosnaíonn díolúine trí vacsaíniú an duine imdhíontaithe ach cosnaíonn sí daoine neamh vacsaínithe sa phobal mar shampla naíonáin atá ró óg le vacsaíniú .
Microsoft	Cosnaíonn díolúine trí vacsaíniú ní hamháin an duine aonair imdhíonta ach cosnaíonn sé daoine neamh-vacsaínithe sa phobal ar nós naíonáin atá ró-óg le vacsaíniú.
Google	Ní amháin go gcosnaíonn díolúine trí vacsaíniú an duine imdhíonta ach cosnaíonn sé daoine gan vacsaíniú sa phobal mar naíonáin atá ró-óg le vacsaíniú a fháil.
(6) Source	Very rarely a persistent measles virus infection can produce subacute sclerosing panencephalitis SSPE a disease in which nerves and brain tissue degenerate progressively and which is more likely to appear if measles infection occurs at a younger age .
Reference	Is annamh is féidir le hionfhabhtú leanúnach víreas bruitíní galar a tháirgeadh fosclerosing panencephalitis SSPE galar ina ngineann néaróga agus fíochán inchinne de réir a chéile agus ar dóichí go dtaispeánfar é má tharlaíonn ionfhabhtú na bruitíní ag aois níos óige.
Our Model	Go han annamh is féidir le hionfhabhtú leanúnach víreas na bruitíní panencephalitis sclerosing subacute SSPE a tháirgeadh galar ina ndéanann néaróga agus fíochán inchinne díghrádú de réir a chéile agus ar dóigh dó a bheith le feiceáil má tharlaíonn ionfhabhtú na bruitíní ag aois níos óige .
Microsoft	Is annamh is féidir le hionfhabhtú leanúnach víreas bruitíní galar a tháirgeadh fosclerosing panencephalitis SSPE galar ina ngineann néaróga agus fíochán inchinne de réir a chéile agus ar dóichí go dtaispeánfar é má tharlaíonn ionfhabhtú na bruitíní ag aois níos óige.
Google	Go han-annamh is féidir le hionfhabhtú víreas leanúnach na bruitíní panencephalitis sclerosing subacute SSPE galar a bhfuil néaróga agus fíochán inchinne ag dul in olcas go comhleanúnach agus atá níos dóchúla le feiceáil má tharlaíonn ionfhabhtú na bruitíní ag aois níos óige.
(7) Source	There may also be trace amounts of other substances used in the manufacturing process such as ovalbumin a protein found in eggs or neomycin an antibiotic .
Reference	D 'fhéadfadh go mbeadh rianta substaintí eile a úsáidtear sa phróiseas táirgeachta ann amhail ubh albaimin próitéin a fhaightear in uibheacha nó neomycinum antaibheathach .
Our Model	D fhéadfadh go mbeadh rianmhéideanna de shubstaintí eile a úsáidtear sa phróiseas monaraíochta mar ovalbumin próitéin a fhaightear in uibheacha nó neomycin antaibheathach .
Microsoft	D'fhéadfadh go mbeadh rianmhéideanna substaintí eile a úsáidtear sa phróiseas monaraíochta, mar shampla próitéin a fhaightear in uibheacha nó neomycin antaibheathach.
Google	D'fhéadfadh go mbeadh riandúile de shubstaintí eile ann a úsáidtear sa phróiseas monaraíochta mar phróitéin ubhchruthach próitéin a fhaightear in uibheacha nó antaibheathach antaibheathach.

Table 6: Outputs of MT engines for various inputs (8) - (9).

Origin	Sentence
(8) Source	Have you checked for recent antibiotic use drug allergies use of immunosuppressive therapy recent hospitalisation or institutionalisation recent travel outside of Europe and microbiology results for the previous months ?
Reference	Ar sheiceáil tú an fhaisnéis seo a leanas maidir leis an othar úsáid antaibheathach le déanaí ailléirgí i leith drugaí úsáid teiripe imdhíon sochtaí ospidéalú nó institiúidiú le déanaí taisteal lasmuigh den Eoraip le déanaí agus torthaí micribhitheolaíocha do na mhí roimhe ?
Our Model	An ndearna tú seiceáil ar úsáid antaibheathach le déanaí ailléirgí drugaí teiripe frith imdhíonachta a úsáid ospidéalú nó institiúidiú le déanaí taisteal le déanaí lasmuigh den Eoraip agus torthaí micribhitheolaíochta don mhí roimhe sin ?
Microsoft	Ar sheiceáil tú le haghaidh ailléirgí drugaí úsáide antaibheathach le déanaí úsáid teiripe imdhíonachta le déanaí ospidéal nó institiúidiú taisteal le déanaí lasmuigh den Eoraip agus torthaí micribhitheolaíochta do na míonna roimhe sin?
Google	An ndearna tú seiceáil le haghaidh úsáid ailléirgí drugaí le húsáid antaibheathach le déanaí ar úsáid teiripe in-imdhíonachta san ospidéal le déanaí nó institiúidiú taisteal le déanaí lasmuigh den Eoraip agus torthaí micribhitheolaíochta do na míonna roimhe sin?
(9) Source	A vaccine stimulates an immune response and a memory of the body to a specific disease without causing the disease .
Reference	Déanann vacsaín freagairt imdhíonachta a spreagadh agus spreagann cuimhne an choirp ar ghalar ar leith gan an galar sin a theacht air .
Our Model	Spreagann vacsaín freagairt imdhíonachta agus cuimhne an choirp ar ghalar ar leith gan an galar a chur faoi deara .
Microsoft	Spreagann vacsaín freagra imdhíonachta agus cuimhne an choirp ar ghalar ar leith gan an galar a chruthú.
Google	Spreagann vacsaín freagairt imdhíonachta agus cuimhne an choirp ar ghalar ar leith gan an galar a chur faoi deara.

4 Online Availability of MT Engine

In the same manner described by Way et al. [28], our chosen model has been deployed for worldwide access.[15] The distributed structure in which the English-Irish NMT engine will be provided is illustrated in Figure 4. Each language pair has a separate GPU server, and the web interface allows users to input text to translate, which is then directed to the appropriate servers. Various modules exist on the server to provide effective processing of input sent to the engine; DNT (do not translate)/TAG/URL, sentence-splitter, tokeniser, character normaliser, truecaser and spellchecker modules currently exist to divide input text into model-readable input. However, the models from this study were built using the MarianNMT [13] platform, which is structured differently to Tensorflow's

[15] `http://covid19-mt.computing.dcu.ie/`

SavedModel format. Our model makes use of its own Sentencepiece models for tokenising and detokenising text, and the model itself is equipped to handle unknown tokens. As a result, we will use a cTranslate2 model format to improve compatibility with the existing server.

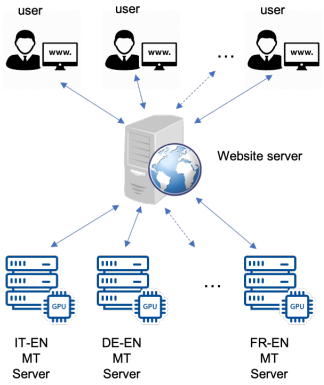

Fig. 4: Webserver configuration for multiple language pair translation [28].

Fig. 5: Web user interface for accessing translation service [28].

5 Conclusion

This paper takes a look at providing COVID-19 domain specific translation for English-Irish. We have led the process all the way from data gathering,

processing and cleaning, to model training and performance comparison with industry leaders, with subsequent deployment for worldwide access.

We have leveraged monolingual data in both English and Irish to improve our model performance through backtranslation techniques. We have tested iterative backtranslation as a method of providing synthetic data for an Irish NMT system, which I have not seen attempted in my research before beginning this project. The scripts used over the course of this project are made available on a GitLab repository.[16]

Our models were trained on freely available data and have proven to be as competitive as – and at times even better than – some of the more popular translation systems, according to widely used evaluation metrics.

Possible future work may include the building and deployment of a corresponding Irish-English MT system, further investigation of model hyperparameters and how they affect MT performance, and the use of additional evaluation metrics such as COMET [23] to rank systems with increased accuracy.

References

[1] Martin Abadi et al. "TensorFlow: A system for large-scale machine learning". In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 2016, pp. 265–283. URL: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

[2] Aivars Berzins et al. *Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe - Why Language Data Matters: ELRC White Paper*. English. 2nd ed. ELRC Consortium, 2019, pp. 94–99. ISBN: 978-3-943853-05-6.

[3] Mihael Arcan et al. "IRIS: English-Irish Machine Translation System". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 566–572. URL: https://www.aclweb.org/anthology/L16-1090.

[4] Marta Bañón et al. "ParaCrawl: Web-Scale Acquisition of Parallel Corpora". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4555–4567. DOI: 10.18653/v1/2020.acl-main.417. URL: https://www.aclweb.org/anthology/2020.acl-main.417.

[5] Arne Defauw et al. "Developing a Neural Machine Translation system for Irish". In: *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. Dublin, Ireland: European Association for Machine Translation, Aug. 2019, pp. 32–38. URL: https://www.aclweb.org/anthology/W19-6806.

[16] https://gitlab.computing.dcu.ie/boylaj25/2021-ca4021-jackboylan

[6] Meghan Dowling, Teresa Lynn, and Andy Way. "Investigating backtranslation for the improvement of English-Irish machine translation". In: *TEANGA, the Journal of the Irish Association for Applied Linguistics* 26 (Nov. 2019). DOI: 10.35903/teanga.v26i0.88.

[7] Meghan Dowling, Andy Way, and Teresa Lynn. "Leveraging backtranslation to improve machine translation for Gaelic languages". In: *Proceedings of the Celtic Language Technology Workshop*. Dublin, Ireland: European Association for Machine Translation, 2019, pp. 58–62. URL: https://www.aclweb.org/anthology/W19-6908.

[8] Meghan Dowling et al. "SMT versus NMT: Preliminary Comparisons for Irish". In: Mar. 2018.

[9] Meghan Dowling et al. "Tapadóir: Developing a Statistical Machine Translation Engine and Associated Resources for Irish". In: 2015.

[10] Rotem Dror et al. "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1383–1392. DOI: 10.18653/v1/P18-1128. URL: https://www.aclweb.org/anthology/P18-1128.

[11] Rohit Gupta et al. *Improving Robustness in Real-World Neural Machine Translation Engines*. 2019. arXiv: 1907.01279 [cs.CL].

[12] Vu Cong Duy Hoang et al. "Iterative Back-Translation for Neural Machine Translation". In: *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 18–24. DOI: 10.18653/v1/W18-2703. URL: https://www.aclweb.org/anthology/W18-2703.

[13] Marcin Junczys-Dowmunt et al. "Marian: Fast Neural Machine Translation in C++". In: Jan. 2018, pp. 116–121. DOI: 10.18653/v1/P18-4020.

[14] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: (2017). arXiv: 1412.6980 [cs.LG].

[15] Guillaume Klein et al. "Efficient and High-Quality Neural Machine Translation with OpenNMT". In: *Proceedings of the Fourth Workshop on Neural Generation and Translation*. Online: Association for Computational Linguistics, July 2020, pp. 211–217. DOI: 10.18653/v1/2020.ngt-1.25. URL: https://www.aclweb.org/anthology/2020.ngt-1.25.

[16] Guillaume Klein et al. "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *CoRR* abs/1701.02810 (2017). arXiv: 1701.02810. URL: http://arxiv.org/abs/1701.02810.

[17] Philipp Koehn and Rebecca Knowles. "Six Challenges for Neural Machine Translation". In: *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, Aug. 2017, pp. 28–39. DOI: 10.18653/v1/W17-3204. URL: https://www.aclweb.org/anthology/W17-3204.

[18] Taku Kudo and John Richardson. "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Process-

ing". In: *CoRR* abs/1808.06226 (2018). arXiv: 1808.06226. URL: http://arxiv.org/abs/1808.06226.

[19] Y. Liu et al. "A Comparable Study on Model Averaging, Ensembling and Reranking in NMT". In: *NLPCC*. 2018.

[20] Kishore Papineni et al. "BLEU: a Method for Automatic Evaluation of Machine Translation". In: (Oct. 2002). DOI: 10.3115/1073083.1073135.

[21] Alberto Poncelas et al. "Investigating Backtranslation in Neural Machine Translation". In: *CoRR* abs/1804.06189 (2018). arXiv: 1804.06189. URL: http://arxiv.org/abs/1804.06189.

[22] Maja Popovic. "chrF: character n-gram F-score for automatic MT evaluation". In: Sept. 2015. DOI: 10.18653/v1/W15-3049.

[23] Ricardo Rei et al. "COMET: A Neural Framework for MT Evaluation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2685–2702. URL: https://www.aclweb.org/anthology/2020.emnlp-main.213.

[24] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Improving Neural Machine Translation Models with Monolingual Data". In: *CoRR* abs/1511.06709 (2015). arXiv: 1511.06709. URL: http://arxiv.org/abs/1511.06709.

[25] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *CoRR* abs/1508.07909 (2015). arXiv: 1508.07909. URL: http://arxiv.org/abs/1508.07909.

[26] Dimitar Shterionov et al. "Human versus automatic quality evaluation of NMT and PBSMT". In: *Machine Translation* 32 (Sept. 2018). DOI: 10.1007/s10590-018-9220-z.

[27] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[28] Andy Way et al. "Rapid Development of Competitive Translation Engines for Access to Multilingual COVID-19 Information". In: *Informatics* 7.2 (June 2020), p. 19. ISSN: 2227-9709. DOI: 10.3390/informatics7020019. URL: http://dx.doi.org/10.3390/informatics7020019.

[29] F. Wilcoxon. "Individual Comparisons by Ranking Methods". In: *Biometrics* 1 (1945), pp. 196–202.