analytics

MDPI

Article

Building Neural Machine Translation Systems for Multilingual Participatory Spaces

Pintu Lohar, Guodong Xie, Daniel Gallagher and Andy Way *

ADAPT Centre, School of Computing, Dublin City University, D09 E432 Dublin, Ireland; pintu.lohar@adaptcentre.ie (P.L.); guodong.xie@adaptcentre.ie (G.X.); daniel.gallagher@adaptcentre.ie (D.G.)
* Correspondence: andy.way@adaptcentre.ie

Abstract: This work presents the development of the translation component in a multistage, multilevel, multimode, multilingual and dynamic deliberative (M4D2) system, built to facilitate automated moderation and translation in the languages of five European countries: Italy, Ireland, Germany, France and Poland. Two main topics were to be addressed in the deliberation process: (i) the environment and climate change; and (ii) the economy and inequality. In this work, we describe the development of neural machine translation (NMT) models for these domains for six European languages: Italian, English (included as the second official language of Ireland), Irish, German, French and Polish. As a result, we generate 30 NMT models, initially baseline systems built using freely available online data, which are then adapted to the domains of interest in the project by (i) filtering the corpora, (ii) tuning the systems with automatically extracted in-domain development datasets and (iii) using corpus concatenation techniques to expand the amount of data available. We compare our results produced by the domain-adapted systems with those produced by Google Translate, and demonstrate that fast, high-quality systems can be produced that facilitate multilingual deliberation in a secure environment.

Keywords: neural machine translation; domain adaptation; parallel data; deliberative democracy; citizens' assemblies

Citation: Lohar, P.; Xie, G.; Gallagher, D.; Way, A. Building Neural Machine Translation Systems for Multilingual Participatory Spaces. *Analytics* **2023**, 2, 393–409. https://doi.org/10.3390/analytics2020022

Academic Editor: Jong-Min Kim

Received: 18 February 2023
Revised: 17 April 2023
Accepted: 24 April 2023
Published: 1 May 2023

1. Introduction

In the European-funded project EUComMeet (https://www.eucommeet.eu/, accessed on 10 December 2022), we need to facilitate multilingual communication between speakers in the context of deliberative democracy. In many European countries, citizens' assemblies have been set up as a response to the challenges currently besetting liberal democracies. These assemblies are participatory spaces created to improve the democratic practice by directly linking citizens with policy makers. Many of the issues currently faced across Europe are pervasive (e.g., immigration, climate change, the Ukraine war, the cost of fuel, etc.), so discussions are taking place in these respective citizens' assemblies, although only in a monolingual context. We aim to permit cross-lingual communication between European citizens on these important topics via machine translation (MT), considerably extending our initial work in Lohar et al. [1].

Sceptics might ask why these participants cannot instead all use English. Despite the widespread claims of English being a lingua franca, this is far from being an actuality today. In any case, in a European context, it flies in the face of our very ideals; *in varietate concordia* ("united in diversity"), the official Latin motto of the EU, illustrates that the many different cultures, traditions and languages in Europe are a hugely positive asset for the continent. In Europe's multilingual setup, all 24 official EU languages are granted equal status by the EU Charter (https://commission.europa.eu/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_en, accessed on 24 November 2022) and the various Treaties of the EU. (https://en.wikipedia.org/wiki/

Treaties_of_the_European_Union, accessed on 12 December 2022) Moreover, the EU is home to 60+ regional and minority languages, which have been protected and promoted under the European Charter for Regional or Minority Languages treaty since 1992 (https://en.m.wikipedia.org/wiki/European_Charter_for_Regional_or_Minority_Languages, accessed on 12 December 2022), in addition to migrant languages and various sign languages, spoken by some 50 million people.

Indeed, the findings of the European Language Equality project (https://european-language-equality.eu/, accessed on 14 December 2022) over the past two years demonstrate a very sorry state of affairs: despite the obvious improvements in language technology since the implementation of methods based on neural networks, language barriers still hamper cross-lingual communication and the free flow of knowledge across borders, and many languages are endangered or on the edge of extinction [2,3]. On a global scale, the situation is far worse, of course, especially for languages that do not have a written tradition [4].

Accordingly, then, translation technology has a vital role to play in addressing these problems, but its capabilities have been hugely overhyped. While in principle, the techniques involved in neural machine translation (NMT) apply to any pair of languages, in practice, demands on data availability restrict these to a small subset of the world's languages, so claims by multinational corporations of "bridging the gap between human and machine translation [quality]" [5] or "human parity" [6] are massively overblown (cf. [7,8] for some of the reasons why).

Given that participants in the deliberative democracy process in different European countries have heretofore been unable to participate in synchronous discussions, one of the aims of the EUComMeet project was to facilitate cross-lingual communication between users by using NMT as a bridge, in two domains of interest across the EU: (i) environment and climate change; and (ii) the economy and inequality.

Large datasets are a prerequisite for high-quality NMT, but these are rarely available for most language pairs and application domains. For the two domains of the project, to the best of our knowledge, no parallel dataset was available, at least for the language pairs with which we are concerned. Accordingly, we began our search for high-quality data in two domains that subsumed those we were interested in, namely economics and the environment. Our aim was to automatically extract tuning and test datasets in these domains with which we could adapt systems trained on generic datasets. We justify this approach in Section 2, while explaining what alternative methods might have been chosen (e.g., building brand new domain-specific engines from scratch, or multilingual models from the outset).

Figure 1 shows the significance of MT systems in multilingual participatory deliberation. The upper part of this figure represents people from different countries wanting to engage in the deliberative democratic process with their peers. They speak to one another in either the economics or environment domain in their own languages. In the likely scenario where they do not understand each other as they belong to different countries with different native languages, it is crucial to translate specific languages to one or more target languages, so that content expressed in those languages will be comprehensible to all participants and not just those for whom the languages are intelligible. For example, if an Italian person is talking to a German and an Irish person, their speech must be translated into both German and Irish or English so that the (German and Irish) listeners can understand what the speaker is saying. The lower part of Figure 1 shows that MT systems lie at the heart of the whole translation process, facilitating translation from one language to another. To facilitate communication between speakers of the six languages, we needed to build 30 ($n(n-1)$) NMT models.

The remainder of this paper is organised as follows: Section 2 summarises the related work in this field. We discuss the overall architecture of MT systems in Section 3, followed by the details of the data collection process in Section 4. We describe our experiments in Section 5, and present the results in Section 6. Finally, we conclude our work, together with some future possible research avenues, in Section 7.

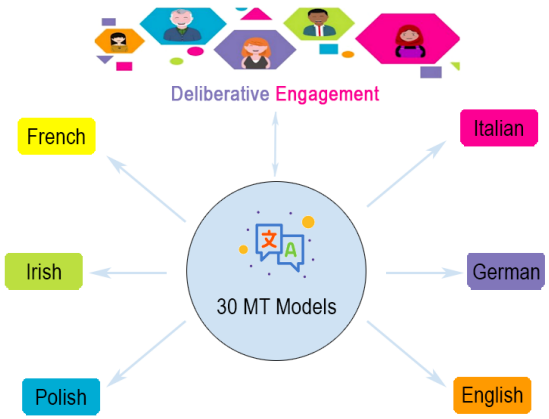

Figure 1. The role of MT in multilingual deliberative discussion.

2. NMT Background

There has been a significant improvement in translation quality in the past few years since the advent of neural MT systems. The earliest NMT systems used encoder–decoder recurrent neural networks (RNNs) [9,10], where the encoder mapped the input string into a hidden representation, which was input to the decoder to generate the target translation. This architecture was improved upon by Bahdanau et al. [11], who introduced the concept of "attention", which allowed the decoder to use the most relevant source-sentence words as contextual clues when deciding how to choose the most appropriate target words(s) for words that were translationally ambiguous in the same sentence.

Following its introduction in Bahdanau et al. [11], the "Transformer" approach of Vaswani et al. [12] relied solely on attention, while at the same time allowing for the use of much simpler feedforward neural networks. To this day, transformers remain the architecture of choice for a range of NLP tasks, as they have achieved state-of-the-art performance across the board, including in translation, where—especially for high-resource language pairs—researchers have claimed similar performance to human-generated translations [13,14].

We do not propose to go into more detail on these different architectures, all of which have been described elsewhere (see also Koehn [15]). However, we are aware that the topic of this paper is likely to attract readers who have no prior exposure to NMT. Accordingly, while we have attempted to make all technical descriptions understandable to nonexperts, for readers for whom the primary literature might be a little foreboding, very good explanations of the technology have been published for those with a less technical background, or who may be newcomers to the field [16,17].

A range of NMT toolkits have been developed that facilitate the development of high-quality translation engines. These include FAIRSEQ (https://github.com/facebookresearch/fairseq accessed on 15 May 2022) [18] and Marian (https://marian-nmt.github.io accessed on 15 May 2022) [19]. Both of these toolkits have many satisfied users, as they have much to commend them. It really comes down to a matter of personal choice, but for the purposes of the experiments in this paper, we instead use OpenNMT (https://opennmt.net accessed on 15 May 2022) [20], an open-source NMT framework that has been widely adopted in the research community and uses the transformer architecture, to build our 30 EUComMeet NMT systems.

To the best of our knowledge, no prior work exists on the development of multilingual deliberative spaces. However, it is well known in MT that large amounts of high-quality domain-specific data are crucial if good translation outputs are to ensue. For example, a parallel biomedical domain corpus [21] can be useful for building an MT system to translate medical texts. However, we are unaware of parallel corpora available for the economics and environment domains. Accordingly, we immediately ruled out the possibility of building individual MT systems for each language direction.

Despite the overly enthusiastic assertion of van der Meer [22] that "massively multilingual MT [is] a single model that can tackle any language pair in the world" (our emphasis; cf. Way [23] for a refutation), it was an option worth considering here. The "massively multilingual" approach of Arivazhagan et al. [24] is a popular choice when the number of languages to be considered is large, but even in cases where only "moderate" amounts of training data are used, the hardware requirements are enormous. Note too that individual MT systems built for specific language pairs have been shown to outperform multilingual models [25,26]. Finally, as our MT systems were to be the intermediary in real-time communication, the models needed to be accurate, but also small enough to ensure rapid translation; users simply could not wait for ostensibly "better" output to ensue if that made communication slow and awkward. We comment more on speed testing in Section 6.3.

Given the imperative that our solution simply had to work, we rejected all the above possibilities for the reasons indicated, and finally opted to tune high-quality baseline engines using smaller domain-specific datasets. For the initial build, we chose Europarl [27], the proceedings of the European Parliament, which was available in all the languages of interest (although in different amounts, cf. Table 1), and is well known in the MT community as a good-quality, general corpus that is sufficiently large for training NMT systems. In addition, given the range of topics dealt with on a daily basis in the European Parliament, it comprises at least some texts from the domains of interest to us, which further justifies its selection.

Table 1. Parallel corpora for English–French.

Corpus Name	Domain	Number of Sentence Pairs in Millions
CCMatrix	Mixed domain	328.6
WikiMatrix	Mixed domain	6.6
UNPC	Mixed domain	22.9
ParaCrawl	Mixed domain	27.6
EUbookshop	Educational	10.8
CCAligned	Mixed domain	15.5
Europarl	Mixed domain	2.05
TED2020	TED talks	0.4
News-Commentary	News	0.2
QED	Educational	1.0

Once these baseline engines were built (cf. Section 5.2), we constructed more advanced systems using additional datasets (cf. Section 4.2), as well as corpus filtering (cf. Section 5.3.1) and domain adaptation techniques (cf. Section 5.3.2), which in all cases caused system performance to improve (cf. Section 6.1) to the point where the majority of our models outperformed Google Translate (cf. Section 6.2).

3. Architecture of Platform for Multilingual Deliberative Democracy

In our previous work on the rapid prototyping of high-performing engines capable of translating multilingual COVID-19 content [28], we adopted Web API in order to make the systems freely available. As that had worked well previously—albeit in asynchronous mode—we employed the same method here to construct the outer interface of our MT platform. Web API is an extensible framework for building HTTP-based services that can be accessed in a range of applications on different platforms such as the web, Windows, mobile, etc. It works more or less the same way as an HTTP web server, except that it sends data as a response instead of HTML. As Web API is based on the HTTP protocol, it can be visited without being subject to firewall restrictions on different local area networks. We built a Web API specification and HTTP request format protocol. Each transaction request is formatted as an HTTP request according to the protocol and is sent to the Web API server of our MT platform. After the transaction has been completed, the Web API sends the results formatted as HTTP response data to the user, as illustrated in Figure 2.

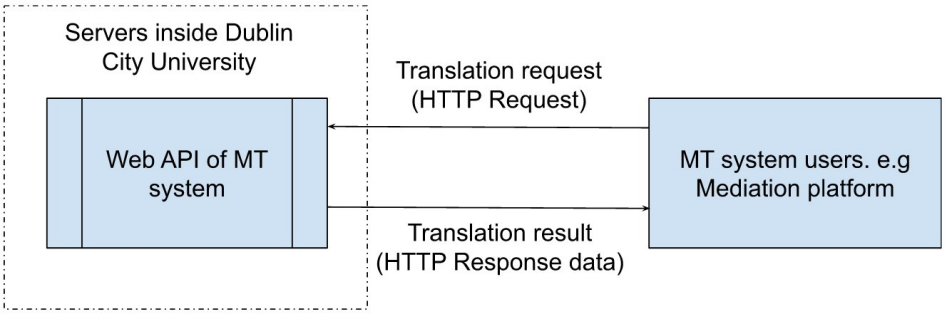

Figure 2. Working process of EUComMeet MT platform Web API.

Our MT platform adopts a two-level server architecture, where Level 1 is the Web API server and Level 2 comprises the translation GPU servers. This is illustrated in Figure 3. Level 1 is the Web API server, which is responsible for accepting HTTP translation requests from the EUComMeet M4D2 platform, where access verification is handled. Here, each user is assigned a unique password-protected username, and to enhance security, we adopt the JSON Web Token (JWT) to verify user access. JWT is an open industry standard used to share information between two entities, usually a client (such as the EUComMeet M4D2 platform, located in Italy) and a server (such as our MT systems, located in Ireland). As the Web API server does not perform translation tasks itself, it can respond to HTTP requests promptly without incurring any delay.

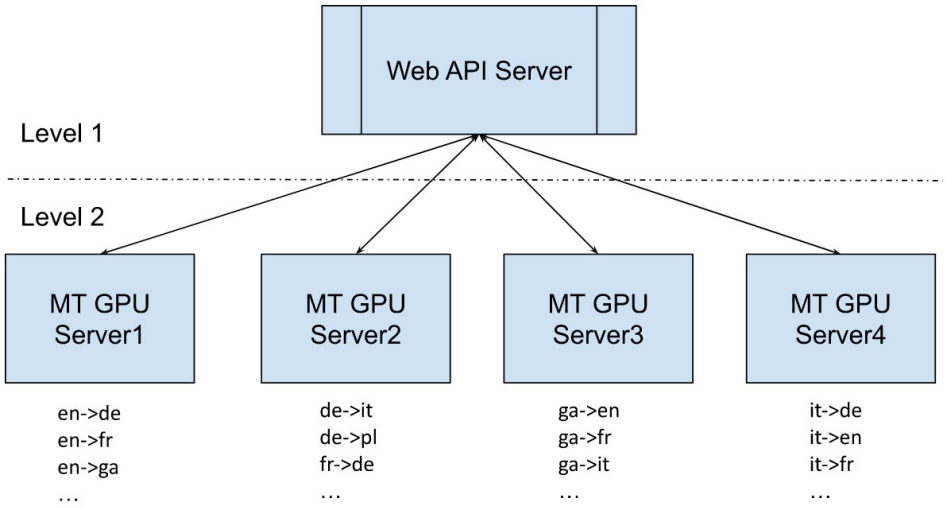

Figure 3. Architecture of EUComMeet MT platform.

Level 1 also performs preprocessing and postprocessing of translation requests. In the preprocessing stage, the text to be translated is first split into individual sentences so that each can be translated in isolation. Then sentences are "normalised" which involves dividing punctuation marks from words, lower-casing all upper-case forms, escaping some special characters, etc. It performs these similar tasks in reverse (i.e., reinserting special characters, truecasing, reattaching punctuation marks to words, etc.) in the postprocessing stage.

Level 2 comprises the actual MT GPU servers, which are responsible for translating the specific source-language text into the requested target language. The GPU servers are all equipped with powerful GPU cards, e.g., GeForce® GTX 1080, GeForce GTX TITAN X, etc. Our MT platform includes 4 GPU servers, on which all 30 MT models are loaded in line with each server's computing power. As these GPU servers are independent of one another, they can work in parallel, which speeds up the translation tasks considerably.

When the Web API receives a translation request, it distributes the translation tasks to each GPU server in parallel and collects the translation results from each one. When all the

translation tasks are complete, the web API packages the results as an HTTP response and sends it back to the user. The chosen two-level architecture ensures the extensibility and efficiency of the whole system. If we find that 4 GPU servers are not enough to support a large number of synchronous translation requests, more GPU servers can be added to the platform without changing the Web API.

4. Data Collection

As already noted, data collection is the most important stage of developing MT systems. While several freely available corpora exist online for training MT systems in many language pairs—especially for major languages such as English, French and German—not all of them are useful when specific tasks are considered. Accordingly, data collection involves the following two substages: (i) corpus survey and (ii) corpus consideration.

4.1. Corpus Survey

An important part of the experimentation phase is exploring what parallel corpora are available online. The following characteristics are crucial for a parallel corpus: (i) the source- and target-language sentences must be good-quality translations of each other; (ii) the corpus should be large enough that it contains at least a few million good-quality sentence pairs; and (iii) the texts should cover the broad domains of interest (in our case, economics and the environment).

A parallel corpus can be considered ideal if it satisfies all of the above requirements, but that is rarely the case. Accordingly, we focused on identifying a target corpus that shared some of these key characteristics. We evaluated a wide range of corpora available on the Internet. OPUS [29], (https://opus.nlpl.eu/ accessed on 6 May 2022) the open parallel corpus, is one of the largest collections of publicly available parallel data from a range of domains for all language pairs of the project. We performed extensive research on a number of parallel corpora available in OPUS.

4.2. Corpus Consideration

For ease of exposition, we consider in Table 1 only the English–French pair to show which corpora we manually inspected.

Table 1 shows the name of the corpora, domains and their sizes in terms of total number of sentence pairs they contain. Despite being an extremely large corpus, we found the "CCMatrix" dataset to be very noisy and not suitable for MT training. Some of the other very large corpora, such as "UNPC" and "ParaCrawl", which contain more than 20 million sentence pairs, are not good enough either.

Although it is not a very large corpus compared to some available resources, Europarl contains a sufficient number of parallel sentences (around 2 million) to build good-quality baseline NMT engines (cf. Section 5.2). In the next phase, we found that some of the smaller corpora, such as "TED2020", "News-Commentary", "QED", etc., are also good-quality corpora, so we concatenated them to the Europarl data and built larger MT models with improved coverage. In addition, we cleaned some of the larger corpora, such as "CCAligned", "WikiMatrix", etc., from which subsets were extracted and concatenated with the already-combined corpora (Section 5.3.3). Our corpus filtering (Section 5.3.1) and domain adaptation (Section 5.3.2) efforts led us to obtain better results over the baseline models in all cases, as demonstrated in Section 6.

5. Experiments

MT system development in EUComMeet comprises three steps: (i) corpus preprocessing, (ii) building baseline MT models and (iii) building advanced MT models. To build both the baseline and domain-adapted MT models, we use OpenNMT with transformer architecture (cf. Section 2). To evaluate the quality of translation, using human experts is preferable, but as translation professionals may be difficult to recruit, developers monitoring the incremental progress of their engines typically use automatic evaluation

metrics instead (cf. Way [30] for more on the advantages and disadvantages of human and automatic evaluation).

We use sacreBLEU [31] to measure BLEU scores [32]—the most widely used automatic metric in MT—to evaluate our MT outputs. Note that BLEU is a precision-based metric, so higher scores are better; in a nutshell, an MT output is compared against a human reference translation (typically held out from the training data) at the level of 1 to 4 grams (individual words up to sets of contiguous words of max. length 4), and the more matches observed, the higher the BLEU score (modulo a brevity penalty, which discourages the output of unduly short translations, such as "The", which otherwise would be an unfair manipulation of the process).

5.1. Corpus Preprocessing

The first step of training an MT model is preprocessing of the dataset. This is crucial to make the dataset as clean as can be for MT system building. The main methods involved in preprocessing are as follows:

- **Filtering out long sentences:** Extremely long sentences were deleted because MT systems generally produce low-quality translations when confronted with lengthy input. If either side of the parallel data contains too many words (100 words is set as the limit), the sentence pair is discarded.
- **Removing blank lines:** Sentence pairs with no content on either side are removed.
- **Removing sentence pairs with odd length ratios:** Sentences with marginally longer or shorter translations when compared to their original sentences were removed because of the probability of them being incorrect translations. The filtering ratio is 1:3 in our case.
- **Removing duplicates:** All duplicate sentence pairs were discarded.
- **Tokenisation:** We break down the sentences into their most basic elements called "tokens" (e.g., transforming "John's" into "John" and "'s"). Tokenisation is particularly relevant because it is the form in which MT models ingest sentences. In practice, most NMT models are fed with subwords as tokens (see next bullet).
- **Byte-pair encoding (BPE):** Restricted vocabularies (e.g., 30 k or 70 k) are typically used for reasons of computational efficiency. Accordingly, many valid words are out-of-vocabulary items (OOVs). However, most OOVs can be broken down into "subwords" already in the restricted vocabulary. We used BPE [33] to try to help the model infer the meaning of OOVs through their similarity to known words. BPE performs subword regularisation by building a vocabulary using corpus statistics. Firstly, it learns the most frequently occurring sequences of characters and then greedily merges them to obtain new text segments.

5.2. Building Baseline MT Models

We built our standard baseline MT systems with the Europarl data using OpenNMT. However, there are no available domain-specific datasets for the areas of interest to the project, not even small datasets usable for tuning and testing. In this early stage of MT system development, we used Europarl as our training dataset and WMT dev and test datasets (https://www.statmt.org/wmt21/translation-task.html accessed on 17 May 2022) for only the high-resource language pairs (English–German, English–French and English–Italian) because the amount of data available for other language pairs (e.g., Italian–Polish, German–Irish, etc.) as either too small or nonexistent. We used a small part of the parallel corpus as (i) devsets for tuning and (ii) test sets for testing for such language pairs. Note that as Europarl had already been used as training data, we could not use it for tuning and testing. Instead, we had to consider some other corpus, such as the TED2020 (https://opus.nlpl.eu/TED2020.php accessed on 11 May 2022) and News-Commentary (https://opus.nlpl.eu/News-Commentary.php accessed on 11 May 2022) corpora. However, for the English–Irish pair, there was neither Europarl for training nor TED2020/News-Commentary for testing. Therefore, we considered a major part of the

the MultiParacrawl (https://opus.nlpl.eu/MultiParaCrawl.php accessed on 12 May 2022) corpus for training and the remainder for tuning and testing, respectively.

Data Statistics

- **Training data:** We used the Europarl corpus as training data for most of the language pairs of the project. For certain language pairs for which Europarl is not available, we use other corpora such as ParaCrawl, MultiParaCrawl, etc. Table 2 shows the statistics of the training data for all language pairs.

Table 2. Statistics of training data for all language pairs.

Language Pair	Corpus Name	#Sentence Pairs
EN–DE	Europarl	1.95M
EN–FR	Europarl	2.05 M
EN–IT	Europarl	1.94 M
EN–PL	Europarl	550 K
EN–GA	ParaCrawl	1.9 M
DE–FR	Europarl	1.94 M
DE–IT	Europarl	1.83 M
DE–PL	Europarl	500 K
DE–GA	MultiParaCrawl	500 K
FR–IT	Europarl	1.94 M
FR–PL	Europarl	521 K
FR–GA	MultiParaCrawl	550 K
IT–PL	Europarl	594 K
IT–GA	MultiParaCrawl	500 K
PL–GA	MultiParaCrawl	350 K

- **Tuning and testing data:** During the development of our baseline MT models, we initially considered only the economics domain to evaluate the quality of the MT output. For most language pairs, the dev and test datasets were available online. For other language pairs such as de–pl, it–ga, etc., we considered either a part of the training data (excluded from the training data *per se* to avoid artificially inflating the evaluation scores) or other corpora such as TED2020. Note that for all language pairs, as was the case with the training data, both the dev and test data belonged to the general or mixed domain. In order to make them suitable for domain-specific tuning and evaluation, we performed a filtering process using relevant key terms so that the filtered dataset was as similar as possible to the economics domain. Accordingly, we manually prepared a list of key terms for both the economics and environment domains, each of which contains more than 100 key terms for all languages of the project, except Irish (We did not need additional lists of key terms for Irish because there was sufficient domain-specific terminology in the parallel data already). Some example key terms for English from both domains can be seen in Table 3. Note that the key terms in this table were produced during this initial stage of experiments. At a later stage, we extracted an improved version of key terms from Wikipedia containing several hundred key terms in each domain. We will discuss this in the next section.

Table 3. Example key terms from economics and environment domains

Economics	Environment
job digitalisation, inequality, consumers, economics, unemployment, investment, globalization, inflation,	pollution, sustainable, climate, deforestation, greenhouse, global warming, atmosphere, ecosystem

Now, we will discuss how the dev and test datasets are filtered using the domain-specific key terms. Consider the example sentences from the WMT test dataset in Table 4.

Table 4. Filtering data with domain-specific key terms.

Text	Key Terms Present	Selection
The new freedoms imported from the West are often synonymous with chaos, as it happened in the 1990s.	None	×
Food: Where European inflation slipped up	inflation	✓
Which role is Putin going to play after March?	None	×
He's tall, strong, very capable, he has a great chest and long arms.	None	×
Consumers also have had to pay significantly more for vegetable oils and dairy products in 2007.	Consumers	✓

Sentences are selected for dev and test sets only if some domain-specific key terms are present. In Table 4, we can see that the second and the fifth sentences contain such key terms (e.g., 'inflation' and 'Consumers'), so these two sentences were selected as part of our dev or test datasets. Note that this is merely a small fragment of the whole datasets; in total, there were several thousands of sentences in the original data, from which several hundreds were selected as they contain at least one domain-specific key term. It is common practice to hold out several hundred sentence pairs for each of the dev and test datasets when available. In our case, the size of the dev and test datasets ranges from 500 to 3000 sentence pairs depending upon the language pair and overall dataset size.

5.3. Building Advanced MT Models

In the previous section, we discussed the development of baseline models with a single corpus. The next step was to build advanced models using other methodologies and tools. We used the following approaches: (i) corpus filtering, (ii) domain adaptation and (iii) corpus combination. Using such additional methods, we hoped to increase performance.

5.3.1. Corpus Filtering

Although we discussed some corpus filtering techniques in Section 5.1, they are standard techniques that cannot filter a corpus based on, for example, semantic similarity. In other words, they cannot discard sentence pairs that are less likely to be translations of each other, or sentence pairs that would be considered to be noisy training data. If such noise exists to a large extent in the training data, it is likely to produce a poor-quality MT model. Therefore, it is crucial to discard less semantically similar sentence pairs and retain only the good-quality ones. Accordingly, we applied a very widely used tool called "LASER" [34] (https://github.com/yannvgn/laserembeddings accessed on 14 July 2022), which is capable of transforming sentences into language-independent vectors. Similar sentences are mapped to close vectors (in terms of cosine distance), regardless of the input language. LASER is very popular among researchers who work on bilingual text mining, as it supports 93 languages, including those of interest to us, and can effectively be used for measuring sentence similarity. Only those sentence pairs that have a similarity score greater than a threshold of 0.6, the value empirically determined by the authors of the original paper to be the best for this task, are maintained in our training corpus.

5.3.2. Domain Adaptation

Next, we further expanded our domain-specific key terms with the help of Wikipedia. Previously, our key terms list consisted of fewer than 200 key terms. In an extension to the previous experiment, we considered all the key terms from Wikipedia for each domain. We refer to these expanded lists as *EcoTerms* (https://en.wikipedia.org/wiki/Glossary_of_economics accessed on 20 August 2022) and *EnvTerms* (https://en.wikipedia.org/wiki/Glossary_of_environmental_science accessed on 20 August 2022). These lists contain a

wide range of key terms, including those mentioned earlier. Table 5 shows some new examples, including the previous terms.

Table 5. Enhanced version of domain-specific key terms (*EcoTerms* and *EnvTerms*).

Domain	Total Number of Key Terms	Examples
Economics (*EcoTerms*)	557	job digitalisation, inequality, consumers, economics, unemployment, investment, globalization, inflation, Accounting reform, Aggregate supply, Experimental economics, Financial astrology, Intertemporal equilibrium …
Environment (*EnvTerms*)	1368	pollution, sustainable, climate, deforestation greenhouse, global warming, atmosphere, ecosystem, Agricultural biodiversity, Biosecurity protocol, Environmental determinism, Ecopsychology, Integrated geography …

We can see from Table 5 that several new advanced terms (highlighted in blue) are added to each domain over and above those common terms previously considered. The enhanced list of domain-specific key terms allows us to generate more relevant domain-related dev and test sets. Previously, we considered only the economics domain on which the baseline model was tuned and tested. In this stage, we consider those sentence pairs that contain key terms from both domains in order to form new dev and test datasets, which we used to tune and test our new models. As the new dev and test data are more relevant than their previous versions, the results of the MT evaluation in this stage are more reliable than our previous evaluation, as will be shown in Section 6, where performance improved for all language pairs.

5.3.3. Corpus Combination

In this approach, we combined multiple corpora to generate larger training sets than those used for baseline MT training. Initially, we filtered the corpora using the LASER tool and then combined them to ensure the quality of the combined corpus. The corpora we considered for combination were *TED2020*, *QED*, *CCAligned*, *WikiMatrix*, *News-Commentary* and *EUBookShop*, all of which were downloaded from OPUS. Note that as stated previously, these corpora are not very suitable for training "as is", so we filtered them using LASER to make them cleaner.

This approach of corpus combination is time-consuming, and is based on several trials over different corpus combinations. This is an ongoing process, and out of a total of 15 language pairs (i.e., ignoring the translation direction, as the training data—but not the dev and test data—are the same for both German-to-English and English-to-German), we have so far performed corpus combination for 8. The remaining seven pairs are EN–GA, DE–GA, FR–PL, FR–GA, IT–PL, IT–GA and PL–GA; note that five out of these seven pairs involve the Irish language (GA). As of now, we prioritise corpus combinations for the other two language pairs (FR–PL and IT–PL); given our local expertise and in-house crawled datasets [35,36], all MT models involving Irish already produce high BLEU scores, and so there is less room for improvement. However, we plan to perform corpus combination for the Irish language once all the other languages are covered.

Table 6 shows the amount of the training data after corpus combination, including corpus filtering for the eight language pairs for which this has been completed. As a comparison, we also provide the size of the training data for the baseline MT systems. Corpus combination and filtering more than doubles the amount of training data for most language pairs, but even in the worst case (DE–FR), the size of the training data increases by 44%.

Table 6. Training data size after corpus combination.

Language Pairs	Size of Training Data in Baseline MT Systems (#Sentence Pairs)	Size of Training Data in Advanced MT Systems (#Sentence Pairs)
DE–EN	1.95 M	6.86 M
FR–EN	2.05 M	3.37 M
IT–EN	1.94 M	4.7 M
PL–EN	550 K	1.17 M
DE–FR	1.94 M	2.8 M
DE–IT	1.83 M	3.24 M
DE–PL	500 K	1.44 M
FR–IT	1.94 M	3.28 M

5.3.4. Architectural Comparison of Baseline vs. Domain-Adapted Systems

We demonstrate the differences between the training processes of the baseline and the advanced MT systems in Figures 4 and 5.

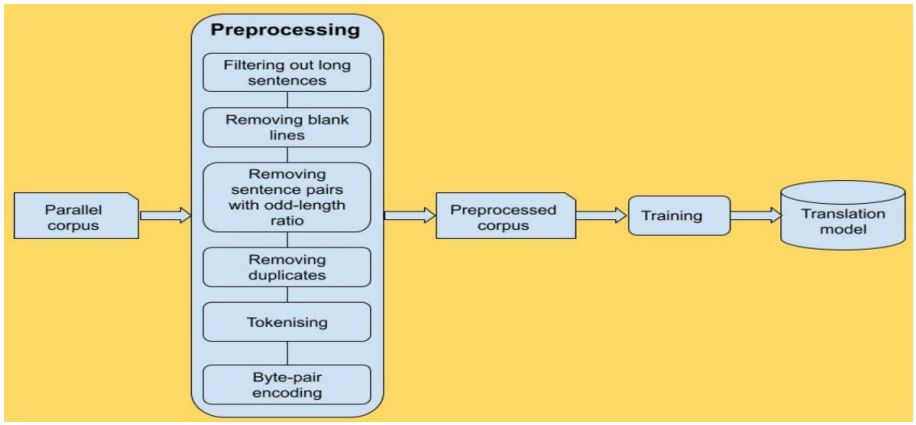

Figure 4. Architecture of baseline MT system training.

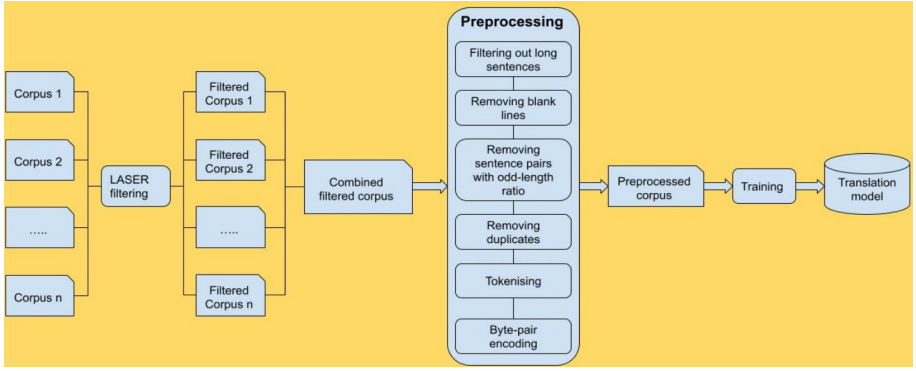

Figure 5. Architecture of domain-adapted MT system training.

The training of the baseline engines shown in Figure 4 comprises a series of simple steps. Firstly, the parallel corpus passes through the preprocessing stage that contains the sequence of simple steps described in Section 5.1, after which it is used to train the baseline models. The training of the advanced models shown in Figure 5 comprises a few additional steps in the initial phase. Firstly, multiple corpora are taken into account. Secondly, each of these corpora is filtered using the LASER tool. Subsequently, the filtered corpora are combined into a single corpus. After that, the process is the same as for the training of the baselines.

6. Results

In this section, we discuss in detail the results obtained by all of the translation models. The performance of the baseline, domain-adapted and Google Translate systems were evaluated on

test data from both the economics and environment domains. We also give some indication of the speed of our models, and initial feedback in user testing of the MT systems to facilitate multilingual discussions in deliberative democracy forums.

6.1. Baseline vs. Domain-Adapted MT System Performance

Firstly, we show the comparison between the baseline and domain-adapted systems in Table 7. The BLEU scores for the domain-adapted models (column 4) all surpass the scores of the corresponding baseline engine (column 3). Columns 5 and 6 show the actual improvement in BLEU score and relative improvement as a percentage, respectively. We highlight all improvements over 10% in bold font in the last column. Two of the largest improvements are seen for German as a target language (from English and French). This is noteworthy, as German is typically regarded as a difficult language to translate, owing to its relatively free word order and rich morphology. The highest improvement is seen for German–Irish, with a huge 23.39% relative improvement, despite this being a difficult language pair.

Table 7. BLEU score comparison for all models. Higher scores are indicative of better-quality translations.

Source Language	Target Language	Baseline System	Domain-Adapted System	BLEU Point Improvement	%Relative Improvement
English	German	23.7	**27.68**	**3.98**	**16.79%**
	French	31.12	**31.58**	0.46	1.47%
	Italian	34.18	**35.58**	1.5	4.38%
	Polish	30.04	**30.25**	0.21	0.69%
	Irish	46.55	**50.99**	4.44	9.53%
German	English	26.67	**27.67**	1.0	3.74%
	French	23.33	**23.8**	0.47	2.01%
	Italian	22.37	**24.59**	2.22	9.92%
	Polish	22.55	**24.39**	1.82	8.07%
	Irish	36.24	**44.72**	**8.48**	**23.39%**
French	German	19.11	**22.85**	**3.74**	**19.57%**
	English	29.39	**31.02**	1.63	5.54%
	Italian	24.56	**26.97**	2.41	9.81%
	Polish	10.01	**10.31**	0.3	2.99%
	Irish	44.69	**45.16**	0.47	1.05%
Italian	German	19.39	**19.50**	0.11	0.56%
	French	26.65	**26.89**	0.24	0.9%
	English	34.62	**35.09**	0.47	1.35%
	Polish	9.65	13.16	3.51	3.63%
	Irish	48.20	**48.42**	0.22	0.45%
Polish	German	24.16	26.13	1.97	8.15%
	French	14.19	**14.29**	0.1	0.7%
	Italian	11.32	**11.66**	0.34	3.0%
	English	38.47	**40.54**	2.07	5.38%
	Irish	38.90	**43.61**	**4.71**	**12.1%**
Irish	German	43.98	**44.79**	0.81	1.84%
	French	47.19	**47.65**	0.46	0.97%
	Italian	48.15	**48.26**	0.11	0.22%
	Polish	37.69	**39.94**	2.25	5.96%
	English	58.34	**62.91**	4.57	7.83%

6.2. Google Translate vs. Domain-Adapted MT System Performance

We compare our domain-adapted systems and Google Translate in Table 8. The last column shows the BLEU-score differences between our domain-adapted systems and Google Translate. The values in red indicate where our MT systems perform worse (preceded by the "−" sign) than their Google Translate counterparts, whereas those in blue show where our systems perform better (preceded by the "+" sign) than Google Translate.

Table 8. BLEU score comparison between domain-adapted models and Google Translate. Higher scores are indicative of better-quality translations.

Source Language	Target Language	Domain-Adapted System	Google Translate	BLEU Point Difference with Google Translate
English	German	27.68	**38.15**	−10.47
	French	**31.58**	28.14	+3.44
	Italian	35.58	**41.55**	−5.97
	Polish	30.25	**36.09**	−5.84
	Irish	**50.99**	44.68	+6.31
German	English	27.67	**42.19**	−15.04
	French	**23.8**	20.89	+2.91
	Italian	**24.59**	24.52	+0.07
	Polish	24.39	**25.39**	−1.0
	Irish	**44.72**	32.27	+12.45
French	German	**22.85**	22.57	+0.28
	English	31.02	**37.38**	−6.36
	Italian	**26.97**	25.2	+1.77
	Polish	10.31	**17.87**	−7.56
	Irish	**45.16**	32.36	+12.8
Italian	German	19.50	**24.43**	−4.93
	French	**26.89**	23.36	+3.53
	English	35.09	**42.72**	−7.63
	Polish	9.65	**16.46**	−6.81
	Irish	**48.42**	34.55	+13.87
Polish	German	26.13	**26.43**	−0.3
	French	14.29	**18.02**	−3.73
	Italian	11.66	**17.76**	−6.1
	English	40.54	**42.16**	−1.62
	Irish	**43.61**	29.93	+13.68
Irish	German	**44.54**	27.74	+16.8
	French	**47.65**	30.29	+17.36
	Italian	**48.26**	33.92	+14.34
	Polish	**39.94**	33.49	+6.45
	English	**62.91**	51.34	+11.57

We can see from this table that out of 30 language pairs, 16 of our domain-adapted systems perform better than Google Translate. It is interesting to note that all models containing Irish as either a source or target language outperform Google Translate. While this may be unexpected in general, it was not a surprise to us; as mentioned in Section 5.3.3, we have access to many good-quality datasets for the Irish language (e.g., in the areas of health [35] and the legal domain [36]) from previous projects, and this considerably improves translation performance compared to the other language pairs, despite not being in the specific areas covered by EUComMeet. These datasets are not in the public domain, so are unavailable to Google Translate as additional training data to improve their engines.

It should be noted that some of our models including other languages such as English–French and German–French also outperform Google Translate. In other cases, the difference in performance is so slight that no human user of the MT systems would be able to detect any difference in performance. As an example, the EUComMeet system obtains a score of 26.13 for Polish-to-German, while Google Translate scores 26.43. This is a difference of only 0.3 BLEU points, or 1.1% relative. In other cases, such as for Italian–English, even though Google Translate achieves a 7.63 higher BLEU score, the EUComMeet engine's score of 35.09 is still relatively high, and indicative of being of value in the intended use-case. By the same token, where performance is low, either for Google Translate or the EUComMeet

model, these engines are unlikely to be effective in user testing in the envisaged use-case. For example, although the Google Translate system for Polish–Italian is 6 BLEU points better, a score of 17.76 is indicative of relatively poor performance. Accordingly, the experience of an Italian user viewing translations from Polish is likely to be quite poor, no matter what system is used for translation.

Given that the EUComMeet engines outperform Google Translate on most language pairs, together with the fact that there are a number of well-known problems associated with the use of Google Translate (including security, cost and customisation), it was agreed by the project consortium to integrate the engines built by EUComMeet into the final deliberative democracy framework. We next describe how translation requests are transmitted to/from the MT systems with fast response times.

6.3. Speed Tests of our MT Systems

In the previous sections, we discussed the performance of our MT systems in terms of quality, as measured by the BLEU score. As mentioned earlier, we needed to not only build systems that could output good-quality translations, but also engines that were fast enough to not hinder the user experience in our use-case of multilingual deliberative democracy.

To give a flavour of the translation speed of our models, over the course of 7 days, the translation latency—the time it takes to send a request to the server and receive back a translation—was measured once per day in seconds. The sentence "This is the EUComMeet project" was sent to the system in each source language and was then translated into the remaining five languages. That is, the English sentence would be sent to the server and the translation in Italian, German, French, Polish and Irish would be received back. In Table 9, we see the average latency time for receiving the translation in five target languages throughout the week. We see that the average remains below one second in each case, a speed that we are confident should facilitate communication in the multilingual deliberative democracy use-case.

Table 9. Average speed of the MT systems for source → all languages (16 December 2022).

Source Language	Latency (s)
English	0.88
Italian	0.53
French	0.85
German	0.96
Polish	0.57
Irish	0.40

6.4. Initial User Reaction

In the three subsections above, we report the accuracy of our systems using BLEU score, and demonstrate performance improvements by domain adaptation, which helps the majority of systems outperform the equivalent Google Translate engine. We also demonstrate that the translation speed is fast, in addition to being accurate.

In initial testing in December 2022, with hundreds of student users, the viability of using the MT systems as an aid for multilingual deliberative democracy was confirmed. There were very few problems with the translations, with even the translations from Polish to English working well. One of the senior moderators noted that the translation from Italian to French worked very well. Overall, for short posts, for language pairs involving English, overall performance was deemed to be good; for language pairs involving Polish, there were some minor discrepancies, and occasionally the translation performance was unreliable; for language pairs involving French, German and Italian, performance was largely good, and at worst acceptable. Irish did not feature in this preliminary user-testing phase, but it will in the large-scale evaluation taking place in May and June 2023; given the relatively good performance of the engines involving the Irish language, we are confident that interactions involving Irish will be successful.

Accordingly, the student testers have been able to conduct their discussions without too many issues, although performance on longer inputs was worse, as expected. A specific lesson learned from the pilot testing was that problems could ensue if the moderator of the multilingual dialogue was unable to speak the languages of all the participants in order to properly explain the process to them. Furthermore, we found that clear instructions should be shared with participants before the event starts, and instructions and guidelines should be easily available (website, reminders, welcome mails, FAQs, etc.) during the deliberation process itself.

As mentioned above, large-scale testing of the system will take place in the near future, but the signs are positive that—for short sentences at least—the MT set-up we have built will sufficiently support multilingual discussions on deliberative democracy for European citizens. If confirmed, this will be a significant step forward to facilitating the joint sharing of views across European countries, the forming of multination policies (as opposed to the current situation whereby recommendations are received from individual countries) via citizens' assemblies and reaffirming the benefits of continued membership of the EU.

7. Conclusions and Future Work

In order to facilitate multilingual participation in deliberative democratic meetings, we have described how the EUComMeet project has built a number of MT systems for all 30 language pairs of the project, for the domains of (i) the environment and climate change and (ii) the economy and inequality. We described how the data underpinning these systems were curated and cleaned, and how improvements were made to the baseline systems. We documented the performance of each MT system built, and compared the quality of each system against Google Translate. Given that the EUComMeet engines outperform Google Translate on most language pairs, together with the fact that there are a number of well-known issues associated with the use of Google Translate, the engines built by EUComMeet have been deployed in the project's deliberative democracy framework. All engines process translation requests with fast response times, and to the general satisfaction of users in initial testing, at least for short texts.

In ongoing work, we plan to conduct a human evaluation of the MT models, which will largely be informed by the large-scale testing due to take place in a few months' time. It will be interesting to discover whether the human evaluation corroborates the results obtained via automatic metrics, and to see whether the generally encouraging results from the initial pilot testing are maintained with a larger set of users. We are currently conducting further investigations into those specific MT models that perform worse than Google Translate, and the gaps are closing. We are encouraged to see that with additional work, our own engines can be tweaked to be competitive with an extremely solidly performing baseline such as Google Translate.

Finally, significant efforts such as those described here are being made in Europe to facilitate effective communication between speakers of different languages. Allowing people to speak their own languages simplifies the process, as users are immediately more comfortable, and coming from different cultural and linguistic backgrounds ensures a richer, more beneficial experience for all, with better decision making and improved outcomes as a result. If this is to be extended to speakers of all European languages—not just those that have ample resources, but languages without a strong written tradition, as well as nonoral languages such as sign languages—then all our languages need to be protected and supported so that they can continue to thrive, so that speakers can operate using their language of choice, as opposed to one that is imposed on them.

Author Contributions: Conceptualization, P.L. and G.X.; Methodology, P.L. and G.X.; Resources, P.L. and G.X.; formal analysis, P.L., G.X. and D.G.; investigation, P.L. and D.G.; writing—review and editing, P.L., G.X., D.G. and A.W.; Supervision, A.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the European Commission under H2020-EU.3.6.—SOCIETAL CHALLENGES—Europe In A Changing World—Inclusive, Innovative And Reflective Societies, grant agreement ID: 959234. This research was also conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Dublin City University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lohar, P.; Xie, G.; Way, A. Developing Machine Translation Engines for Multilingual Participatory Spaces. In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, Ghent, Belgium, 1–3 June 2022; European Association for Machine Translation: Paris, France, 2022; pp. 359–360.
2. Gaspari, F.; Gallagher, O.; Rehm, G.; Giagkou, M.; Piperidis, S.; Dunne, J.; Way, A. Introducing the Digital Language Equality Metric: Technological Factors. In Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; European Language Resources Association: Paris, France, 2022; pp. 1–12.
3. Grützner-Zahn, A.; Rehm, G. Introducing the Digital Language Equality Metric: Contextual Factors. In Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 13–26.
4. Bird, S. Local Languages, Third Spaces, and other High-Resource Scenarios. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; pp. 7817–7829. [CrossRef]
5. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. In Proceedings of the Transactions of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 5, pp. 339–351.
6. Hassan, H.; Aue, A.; Chen, C.; Chowdhary, V.; Clark, J.; Federmann, C.; Huang, X.; Junczys-Dowmunt, M.; Lewis, W.; Li, M.; et al. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv* **2018**, arXiv:1803.05567.
7. Läubli, S.; Sennrich, R.; Volk, M. Has Machine Translation Achieved Human Parity? A Case for Document-Level Evaluation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, 31 October–4 November 2018; pp. 4791–4796.
8. Toral, A.; Castilho, S.; Hu, K.; Way, A. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In Proceedings of the Third Conference on Machine Translation (WMT18), Brussels, Blgium, 31 October–1 November 2018; pp. 113–123.
9. Cho, K.; van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In Proceedings of the SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014; pp. 103–111. [CrossRef]
10. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27, 9p.
11. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
13. Barrault, L.; Bojar, O.; Costa-jussà, M.R.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Koehn, P.; Malmasi, S.; et al. Findings of the 2019 Conference on Machine Translation (WMT19). In Proceedings of the Fourth Conference on Machine Translation, Florence, Italy, 1–2 August 2019; pp. 1–61.
14. Bawden, R.; Bretonnel Cohen, K.; Grozea, C.; Jimeno Yepes, A.; Kittner, M.; Krallinger, M.; Mah, N.; Neveol, A.; Neves, M.; Soares, F.; et al. Findings of the WMT 2019 Biomedical Translation Shared Task: Evaluation for MEDLINE Abstracts and Biomedical Terminologies. In Proceedings of the Fourth Conference on Machine Translation, Florence, Italy, 1–2 August 2019; pp. 29–53.
15. Koehn, P. *Neural Machine Translation*; Cambridge University Press: New York, NY, USA, 2020.
16. Forcada, M.L. Making sense of neural machine translation. *Transl. Spaces* **2017**, *6*, 291–309. [CrossRef]
17. Way, A. Machine translation: Where are we at today? In *The Bloomsbury Companion to Language Industry Studies*; Angelone, E., Massey, G., Ehrensberger-Dow, M., Eds.; Bloomsbury: London, UK, 2019; pp. 311–332.
18. Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; Auli, M. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, MN, USA, 2–7 June 2019; pp. 48–53. [CrossRef]

19. Junczys-Dowmunt, M.; Grundkiewicz, R.; Dwojak, T.; Hoang, H.; Heafield, K.; Neckermann, T.; Seide, F.; Germann, U.; Aji, A.F.; Bogoychev, N.; et al. Marian: Fast Neural Machine Translation in C++. In Proceedings of the ACL 2018, System Demonstrations, Melbourne, VI, Australia, 15–20 July 2018; pp. 116–121. [CrossRef]
20. Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In Proceedings of the ACL 2017, System Demonstrations, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 67–72.
21. Névéol, A.; Jimeno Yepes, A.; Neves, M.; Verspoor, K. Parallel Corpora for the Biomedical Domain. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018; pp. 291–296.
22. van der Meer, J. Multilingual Magazine. Translation Economics of the 2020s: A Journey into the Future of the Translation Industry in Eight Episodes. 2021. Available online: https://multilingual.com/issues/july-august-2021/translation-economics-of-the-2020s/ (accessed on 10 September 2022).
23. Way, A. The Future of Translation Technologies in Society. In *Routledge Handbook of Translation Technology and Society*; Baumgarten, S., Tieber, M., Eds.; Routledge: Abingdon, UK, 2023.
24. Arivazhagan, N.; Bapna, A.; Firat, O.; Aharoni, R.; Johnson, M.; Macherey, W. The Missing Ingredient in Zero-Shot Neural Machine Translation. *arXiv* **2019**, arXiv:1903.07091.
25. Lu, Y.; Keung, P.; Ladhak, F.; Bhardwaj, V.; Zhang, S.; Sun, J. A neural interlingua for multilingual machine translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, Brussels, Belgium, 31 October–1 November 2018; pp. 84–92.
26. Johnson, M.; Schuster, M.; Le, Q.V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 339–351. [CrossRef]
27. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In Proceedings of the Machine Translation Summit X: Papers, Phuket, Thailand, 13–15 September 2005; pp. 79–86.
28. Way, A.; Haque, R.; Xie, G.; Gaspari, F.; Popović, M.; Poncelas, A. Rapid Development of Competitive Translation Engines for Access to Multilingual COVID-19 Information. *Informatics* **2020**, *7*, 19. [CrossRef]
29. Tiedemann, J. OPUS—parallel corpora for everyone. In Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products, Riga, Latvia, 30 May–1 June 2016.
30. Way, A. Quality Expectations of Machine Translation. In *Translation Quality Assessment: From Principles to Practice*; Moorkens, J., Castilho, S., Gaspari, F., Doherty, S., Eds.; Springer: Cham, Switzerland, 2018; pp. 159–178.
31. Post, M. A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels, 31 October–1 November 2018; pp. 186–191.
32. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
33. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 1715–1725. [CrossRef]
34. Artetxe, M.; Schwenk, H. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 597–610. [CrossRef]
35. Lankford, S.; Afli, H.; Ní Loinsigh, Ó.; Way, A. gaHealth: An English–Irish Bilingual Corpus of Health Data. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 6753–6758.
36. Bago, P.; Castilho, S.; Celeste, E.; Dunne, J.; Gaspari, F.; Gíslason, N.R.; Kåsen, A.; Klubička, F.; Kristmannsson, G.; McHugh, H.; et al. Sharing high-quality language resources in the legal domain to develop neural machine translation for under-resourced European languages. *Rev. Lleng. Dret J. Lang. Law* **2022**, *78*, 9–34.