NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

PROGRAM OF POSTGRADUATE STUDIES
DATA SCIENCE AND INFORMATION TECHNOLOGIES

SPECIALIZATION
BIG DATA AND ARTIFICIAL INTELLIGENCE

Master's Thesis

Building End-to-End Neural Machine Translation Systems for Crisis Scenarios: The Case of COVID-19

Dimitrios G. Roussis

ATHENS

NOVEMBER 2022

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΕΙΔΙΚΕΥΣΗ
ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Διπλωματική Εργασία

Κατασκευή Ολοκληρωμένων Συστημάτων Νευρωνικής Μηχανικής Μετάφρασης για Καταστάσεις Κρίσεων: Η Περίπτωση του COVID-19

Δημήτριος Γ. Ρούσσης

ΑΘΗΝΑ

ΝΟΕΜΒΡΙΟΣ 2022

Master's Thesis

Building End-to-End Neural Machine Translation Systems for Crisis Scenarios: The Case of COVID-19

Dimitrios G. Roussis

S.N.: DS1190017

SUPERVISOR: **Vassilis Katsouros,** Research Director, "Athena" Research and Innovation Center

EXAMINATION COMMITTEE:

Vassilis Papavassiliou, Research Associate, "Athena" Research and Innovation Center
Sokratis Sofianopoulos, Research Associate, "Athena" Research and Innovation Center
Vassilis Katsouros, Research Director, "Athena" Research and Innovation Center

November 2022

Διπλωματική Εργασία

Κατασκευή Ολοκληρωμένων Συστημάτων Νευρωνικής Μηχανικής Μετάφρασης για Καταστάσεις Κρίσεων: Η Περίπτωση του COVID-19

Δημήτριος Γ. Ρούσσης

Α.Μ.: DS1190017

ΕΠΙΒΛΕΠΩΝ: **Βασίλης Κατσούρος,** Διευθυντής Ερευνών, Ερευνητικό Κέντρο «Αθηνά»

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Βασίλης Παπαβασιλείου,** Συνεργαζόμενος Ερευνητής, Ερευνητικό Κέντρο «Αθηνά»
Σωκράτης Σοφιανόπουλος, Συνεργαζόμενος Ερευνητής, Ερευνητικό Κέντρο «Αθηνά»
Βασίλης Κατσούρος, Διευθυντής Ερευνών, Ερευνητικό Κέντρο «Αθηνά»

Νοέμβριος 2022

ABSTRACT

Machine Translation is a crucial task of Natural Language Processing, as it aims to provide a fast and automatic way of translating various types of texts. In recent years, the emergence of Neural Machine Translation and the compilation of large-scale parallel corpora have led to significant improvements in translation quality. However, translation models are not necessarily suited for all domains and, thus, there has been significant research on domain adaptation of Neural Machine Translation Systems, i.e., on how to best improve the translation quality of an existing system for a specific topic or genre.

Crisis Machine Translation is a special case of Domain Adaptation which is concerned with the rapid adaptation of an existing Machine Translation system for a crisis scenario, as the integration of such a system in a rapid response infrastructure can accelerate the speed of decision making and relief provision. The COVID-19 pandemic proved to be a prolonged and global crisis with large gaps in transparent, timely, and effective communication; it was also marked by misinformation, conspiracy theories, and significant restrictions on press freedom. Further research on Crisis Machine Translation could play an important role in better responding to future similar crises.

In this thesis, we focus on the case of the COVID-19 pandemic and the English-Greek translation direction, while we also create two domain-specific multilingual parallel corpora; one which is related to COVID-19 and one which has been gathered from the abstracts of academic theses and dissertations.

First, we describe the methodologies of acquiring new domain-specific parallel corpora and generating synthetic data which are combined with existing parallel data so as to adapt an existing system to the domain of COVID-19. This process includes data filtering, pre-processing, and selection pipelines, which are also described in detail.

Afterwards, we conduct experiments on different fine-tuning strategies for a simulated crisis scenario in which varying amounts of related data become available as time progresses. We are also concerned with the phenomenon of “catastrophic forgetting”, i.e., the degradation of system performance on general texts.

Lastly, we construct an end-to-end Neural Machine Translation system which is specialized in translating COVID-19 related English texts into Greek. In order to assess its performance across different domains and determine its strengths and weaknesses, we conduct an extended evaluation with eight test sets (half of them have been specifically created for this thesis) and other publicly available models and commercial translation services.

SUBJECT AREA: Neural Machine Translation

KEYWORDS: COVID-19, Domain Adaptation, Crisis Machine Translation, Multilingual Corpora Acquisition, Transformers

ΠΕΡΙΛΗΨΗ

Η Μηχανική Μετάφραση είναι ένα σημαντικό κομμάτι της Επεξεργασίας Φυσικής Γλώσσας, καθώς στοχεύει στην γρήγορη και αυτόματη μετάφραση διαφόρων ειδών κειμένων. Τα τελευταία χρόνια, η επικράτηση της Νευρωνικής Μηχανικής Μετάφρασης και η δημιουργία μεγάλων παράλληλων συλλογών κειμένων έχει οδηγήσει σε σημαντική βελτίωση της ποιότητας μετάφρασης. Ωστόσο, τα μεταφραστικά μοντέλα δεν είναι απαραιτήτως κατάλληλα για όλους τους τομείς κειμένων κι αυτό έχει οδηγήσει σε διάφορες έρευνες σχετικές με την προσαρμογή υπαρχόντων συστημάτων Μηχανικής Μετάφρασης σε διάφορους γνωστικούς τομείς κειμένων, δηλ. στο πώς να βελτιωθεί καλύτερα η ποιότητα μετάφρασης για μία συγκεκριμένη θεματική ή είδος κειμένων.

Η Μηχανική Μετάφραση για Καταστάσεις Κρίσεων είναι μία ιδιαίτερη εφαρμογή της εξειδίκευσης συστημάτων σε θεματικούς τομείς, η οποία ασχολείται με την γρήγορη εξειδίκευση ενός υπάρχοντος συστήματος Μηχανικής Μετάφρασης για μία κατάσταση κρίσης, καθώς η ενσωμάτωση ενός τέτοιου συστήματος σε μία υποδομή ταχείας απόκρισης μπορεί να επιταχύνει την παροχή βοήθειας και την λήψη αποφάσεων. Η πανδημία του COVID-19 αποδείχτηκε μία κρίση μεγάλης διάρκειας και διεθνούς χαρακτήρα στην οποία παρουσιάστηκαν μεγάλα κενά στην διαφανή, έγκαιρη και αποτελεσματική επικοινωνία με το κοινό, ενώ σημαδεύτηκε από παραπληροφόρηση, θεωρίες συνομωσίας και σημαντικούς περιορισμούς στην ελευθερία του Τύπου. Περαιτέρω έρευνα στο πεδίο της Μηχανικής Μετάφρασης για Καταστάσεις Κρίσεων θα μπορούσε να συνδράμει σημαντικά στην αντιμετώπιση παρόμοιων μελλοντικών κρίσεων.

Η παρούσα διπλωματική εργασία εστιάζει στην περίπτωση της πανδημίας του COVID-19 και στην μετάφραση αγγλικών κειμένων στα ελληνικά, ενώ επίσης κατασκευάζονται δύο εξειδικευμένα πολυγλωσσικά παράλληλα σώματα κειμένων. Το ένα σχετίζεται με τον COVID-19 και το άλλο προέρχεται από περιλήψεις ακαδημαϊκών εργασιών και διατριβών.

Στην αρχή περιγράφουμε την διαδικασία συλλογής καινούριων παράλληλων σωμάτων κειμένων για συγκεκριμένους τομείς και την δημιουργία συνθετικών δεδομένων. Αυτά τα δεδομένα συνδυάζονται με υπάρχοντα παράλληλα δεδομένα ώστε να εξειδικεύσουν ένα υπάρχον σύστημα για τον COVID-19. Η διαδικασία αυτή περιλαμβάνει επίσης το φιλτράρισμα, την προεπεξεργασία και την επιλογή κατάλληλων δεδομένων, τα οποία παρουσιάζονται αναλυτικώς.

Έπειτα, κάνουμε πειράματα πάνω σε διαφορετικές στρατηγικές εξειδίκευσης υπαρχόντων συστημάτων Μηχανικής Μετάφρασης για μία προσομοιωμένη κατάσταση κρίσης όπου τα σχετικά δεδομένα αυξάνονται με την πάροδο του χρόνου. Μας ενδιαφέρει επίσης το φαινόμενο “catastrophic forgetting” στο οποίο παρουσιάζεται μείωση της ποιότητας μετάφρασης σε κείμενα γενικού περιεχομένου.

Τέλος, κατασκευάζουμε ένα ολοκληρωμένο σύστημα Νευρωνικής Μηχανικής Μετάφρασης το οποίο είναι εξειδικευμένο στην μετάφραση αγγλικών κειμένων σχετικά με τον COVID-19 στα ελληνικά. Αξιολογούμε διεξοδικά την απόδοσή του σε διαφορετικά είδη κειμένων ώστε να βρούμε τα δυνατά και αδύνατα σημεία του, κάνοντας χρήση οκτώ εξειδικευμένων δοκιμασιών (εκ των οποίων τα τέσσερα δημιουργήθηκαν για την παρούσα διπλωματική) και άλλων διαθέσιμων μοντέλων και υπηρεσιών μετάφρασης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Νευρωνική Μηχανική Μετάφραση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: COVID-19, Προσαρμογή Γνωστικού Αντικειμένου, Μηχανική Μετάφραση για Καταστάσεις Κρίσεων, Συλλογή Πολυγλωσσικών Σωμάτων Κειμένων, Transformers

ACKNOWLEDGEMENTS

This work has been conducted at the Institute of Language and Speech Processing at the “Athena” Research and Innovation Center. It has been supported by the European Language Resource Coordination (ELRC), a service contract under the European Commision’s Connecting Europe Facility SMART 2019/1083 program, as well as the European Language Grid (ELG), a service contract under the Horizon 2020 Information and Communications Technology (H2020-ICT) 29a program.

I want to express my deep appreciation to Vassilis Papavassiliou for being a true and understanding mentor, to Stelios Piperidis for believing in me from day one, and to Sokratis Sofianopoulos as well as all the people at the Institute of Language and Speech Processing which have created -and continue to maintain- a friendly and caring environment.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In this chapter we provide a general introduction to all the topics which concern us, such as the COVID-19 pandemic, Machine Translation (MT), and the importance of Crisis Machine Translation (Crisis MT) in crises scenarios. Furthermore, we give an overview of the thesis, a review of related work, as well as several remarks on the philosophy behind Neural Machine Translation (NMT) and domain adaptation.

1.1 Introductory Notes

In December 2019, the first cases of the COVID-19 disease caused by the SARS-CoV-2 virus were identified in Wuhan, Hubei, China. On 11 March 2020, the WHO declared the novel coronavirus outbreak a pandemic and shortly afterwards, it started affecting the whole world, country by country, to a smaller or larger extent. Thus, COVID-19 unfolded as the catalyst of global and simultaneous changes in everyday life and public discourse, as the threat of collapsed healthcare systems by the pandemic led most countries to impose restrictive measures. These measures, in turn, had a tremendous impact on the global economy, healthcare systems, national policies, working conditions, individual freedoms, journalism, travel, and movement.

Due to the pandemic, we became increasingly dependent on mass media, social media, news websites, online communication services, and, in general, anything that involved the internet and the “digital”. Thus, of COVID-19 related textual information was generated in abundance across many languages and platforms; this motivated researchers to gather and structure data in order to facilitate the use of statistical methods and AI (Artificial Intelligence) for a plethora of applications [1]. The emergence of multidisciplinary initiatives, such as the creation of CORD-19 (Covid-19 Open Research Dataset) and the shared tasks built around it, illustrated how data scientists, biomedical experts, and policy makers can connect and cooperate in the fight against COVID-19 or other future similar crises [2].

Machine Translation is one of the most important tasks of NLP (Natural Language Processing) and the -relatively recent- use of Neural Machine Translation (NMT) has led to significant improvements for many language pairs. Most NMT models perform well on general texts across various domains, although still underperform on specialized texts. Domain adaptation is a family of methods aimed at improving translation quality of an existing system for a specific domain, i.e., an area of knowledge, a specific topic, or a genre with particular syntactic and stylistic properties. Yet even the ones that are adapted to address specific domains, such as legal or biomedical texts, often require benchmarking and continuous improvement. The most obvious example, which is also connected with this thesis, concerns the translation of COVID-19 related news which generic NMT systems trained with news from earlier dates could not easily address [3].

Thus, in this thesis, we aim towards the construction of an end-to-end NMT (Neural Machine Translation) system which can handle COVID-19 related information. In general terms, the tasks which we focused on are the following:

- Acquiring domain-specific parallel data of high quality from various sources. The methodology described in this thesis has led to the construction of several large multilingual corpora (see Section 3.2).

- Applying pre-processing, filtering and cleaning pipelines to pre-existing and newly acquired parallel data, as well as synthetic data generated from monolingual sentences.
- Training Transformer models from scratch using various configurations and investigating the specifics of data augmentation and domain adaptation strategies in order to improve their performance on COVID-19 related texts, while also maintaining their performance on general texts.
- Leveraging additional popular methods and techniques to construct an end-to-end NMT system and developing an extensive evaluation framework in order to compare its performance with other systems and popular commercial translation services.

In particular, we focus on the English to Greek (EN→EL) translation direction and on COVID-19 news and information, as well as other biomedical texts. We investigate different strategies on how to best respond to a scenario in which time is in shortage and the rapid construction of an automatic translation system for domain-specific information (or even a specific language pair) is required. The experiments that we perform concern a crucial element of Crisis Machine Translation (Crisis MT) [4].

Nevertheless, the insights gathered throughout the thesis do not only concern COVID-19, but also other types of -possible future- crises. Thus, various techniques and methods were subsequently used to construct NMT systems which could address needs that emerged from the Ukrainian crisis after its invasion by the Russian Federation in February 2022. This also led to the submission of two NMT systems at the General Machine Translation shared task of the Seventh Conference on Machine Translation (WMT22) for the English to Ukrainian and Ukrainian to English translation directions [5].

Furthermore, it is important to note that the data acquisition process described in this thesis led to the creation and publication of two specialized corpora: (a) SciPar, which is collection of parallel corpora from scientific abstracts [6], and (b) a COVID-19 parallel news resource [7] which was created using metadata from MediSys [8]. The latter was partly used, along with other parallel and monolingual corpora described in this thesis (see Section 3.2.3), in the COVID-19 MLIA-Eval initiative, an initiative focused on facilitating access to COVID-19 related multilingual information [9, 10].

The rest of this thesis is organized as follows:

- The rest of **Chapter 1** provides a review of related work, examines the importance of Crisis MT and its connection with the pandemic, and discusses some useful philosophical considerations on the task of machine translation.
- **Chapter 2** briefly provides background information on Neural Machine Translation, the Transformer architecture used in this thesis, as well as Domain Adaptation. It also outlines some important elements of Crisis MT systems and connects them with the case of the COVID-19 pandemic.
- **Chapter 3** provides information about the existing and newly created datasets that were used in this thesis, as well as the methods and tools utilized in order to acquire parallel data and generate synthetic data. Additionally, it describes the filtering methods and domain categorization that was used on the data.
- **Chapter 4** describes the automatic metrics, as well as the existing -and newly constructed- test sets that were used in the evaluation of (a) the experiments involving domain adaptation and (b) the development of an end-to-end Neural Machine Translation system.

- **Chapter 5** presents the experimentation framework, the system configuration used, and the results of the experiments related to the rapid domain adaptation of existing systems for crisis scenarios.
- **Chapter 6** compares the end-to-end Neural Machine Translation system that was created in this thesis with other openly available systems and commercial translation services.
- **Chapter 7** concludes and summarizes the thesis and discusses potential future research directions.

1.2 Review of Related Work

As mentioned earlier, the pandemic ignited interest in the AI community on many aspects related to our response to it. In this section we describe several benchmarks, initiatives, corpora, and translation systems which emerged during the COVID-19 crisis.

For example, the CORD-19 (Covid-19 Open Research Dataset) is a public resource of scientific papers related to COVID-19 which was created with the aim of facilitating AI applications, such as information retrieval, information extraction, knowledge graphs, question answering, etc. [2].

Several resources more closely related to MT, include the TICO-19 (Translation Initiative for COvid-19) project which published COVID-19 related resources such as benchmarks, translation memories, and terminologies in several language pairs, especially focusing on low-resource languages [11]. The ParaCrawl project utilized a glossary from TICO-19 as well as existing parallel corpora to create a synthesized corpus related to the COVID-19 domain [12]. Additionally, TAUS launched the Corona Crisis Corpus project which released parallel data related to virology, epidemiology, medicine, and healthcare. SYSTRAN contributed to this project with several Corona Crisis Translation Models [13].

There are several shared tasks aimed at assessing the efficacy of NMT models for specialized texts, such as the Biomedical Translation [14] and Machine Translation using Terminologies [15] shared tasks organized by the Conference on Machine Translation (WMT). Specifically related to COVID-19, the 4th LoResMT (Workshop on Technologies for MT of Low Resource Languages) organized shared tasks for COVID-19 texts and sign language and focused on low resource languages, such as Irish, Marathi and Taiwanese Sign Language [16]. The COVID-19 MLIA Eval initiative, which we mentioned earlier, organized shared tasks on (a) Information Extraction, (b) Multilingual Semantic Search, and (c) Machine Translation for COVID-19 information [9, 10].

Finally, there have been several applications and publications which constructed or adapted MT translation systems specifically for COVID-19 related texts. For instance, in [17], an English monolingual corpus is used to adapt an existing NMT system for COVID-19 texts, while in [18] there are experiments on factored NMT (i.e., use of additional linguistic features to augment data) for COVID-19 related texts with five low resource language pairs. More closely related to this thesis is [19], in which eight NMT systems were rapidly developed (and compared with other commercial translation services) in order to facilitate access to multilingual COVID-19 information.

1.3 The Importance of Crisis Machine Translation

MT is a useful tool for emergency situations, as it can been used to accelerate the speed in which relief can be provided; for example, MT had been embedded in the rapid response infrastructure during the Haiti earthquake in 2010 [4], while other notable uses of MT

technology for crisis situations include the efforts by the Translators without Borders and the INTERACT (INTERnAtional network on Crisis Translation) project funded by the EU [11].

The COVID-19 pandemic, however, proved to be a prolonged and global crisis resulting in millions of deaths. The lack of transparent, timely, and effective communication by authorities, as well as the tendency of each state to react only after local transmission was underway, led to slow and ineffective responses [20] and highlighted the need to facilitate access to critical, high-quality information for many stakeholders: healthcare professionals, decision makers, vulnerable or marginalized groups (which may not share a common language with medical personnel), the oldest parts of the population, etc. [7].

The public response to the pandemic was highly mixed; constructive behaviors and adherence to official announcements coexisted with panic buying, stigmatization of COVID-19 patients and Asian communities, and outright denial of the importance of vaccinations [20]. One of the reasons for this, was the “infodemic” taking place in parallel. Overwhelming amounts of misinformation and conspiracy theories (propagated even by public figures) mainly in social media and digital platforms further complicated the task of discriminating between valid and non-valid information [21]. Moreover, the importance and use of digital media became even greater due to the restrictive measures taken to combat the COVID-19 pandemic [22]. On top of all these, many states -even democratic ones- pushed for restrictions on press freedom, with journalists facing multifaceted threats such as prohibitions to attend press briefings, confiscation or destruction of equipment, smear campaigns, digital surveillance, massive cuts in wages, self-censorship due to strict legal measures, etc. [23].

All the above, lead us to consider that there are lessons to be learned from the pandemic and that there is a need for structural changes in informational flows during crisis scenarios. In a world where crises tend to become the norm, the role of Crisis MT becomes even greater as it can provide tools for decision making and the dissemination of high-quality information which, coupled with effective leadership and targeted interventions, could prove vital in mitigating the effects of future crises, especially for the most vulnerable parts of the population. [21]. Furthermore, we believe that Crisis MT could further facilitate rapid access to translated “raw” information from intergovernmental agencies, foreign national authorities, and field experts (publishing mainly in English) and thus, empower journalists to cover certain issues with less degrees of censorship (and self-censorship) and hold public figures accountable for the dissemination of misinformation and/or hateful speech.

1.4 Philosophy of Machine Translation

Translating a text from one language to another is an inherently difficult, open, and vague task. Although the usual goal of translation is to preserve the semantics (i.e., the meaning) of a text across different languages, we make different stylistic choices when, for example, we translate a novel and aim to better immerse the reader (than strictly follow the original author’s prose) compared to when we translate a scientific text and aim at using the correct scientific terminology; thus, having less stylistic liberties.

Nonetheless, the philosopher Willard Van Orman Quine provided arguments to show that there might be different ways of translating a sentence of source language L_{src} to target language L_{trg} which are equally correct and not merely stylistic choices. His famous thought experiment involved a native speaker uttering the word “gavagai” and pointing to a rabbit, and an English-speaking linguist realizing that this utterance may actually mean anything from “rabbit”, to “undetached rabbit parts”, or even “rabbit stages”, etc., since she/he is not exactly certain what the native speaker is referring to [24]. The argument aims to show that an unknown language may use an altogether different conceptual scheme than ours and

the practical takeaway is that a proxy function is actually required to map a specific meaning in a specific context between languages L_{src} and L_{trg} [25]. This could also be interpreted as the basic idea behind NMT and a possible explanation of their significant improvements over past statistical methods (such as Statistical Machine Translation). The improvements have also been driven by the compilation of large parallel corpora, i.e., many real cases of contextualized mappings between two languages.

Another philosopher, Ludwig Wittgenstein, described language as an inherently public activity comprising innumerable "language games" which are connected with a network of similarities, or "family resemblances" [26], i.e., by overlapping similarities and not with a common feature. One may utter the word "Water!" in many "language games", as an exclamation, an order, a request, or an answer to a question; common human behavior is the point of reference which enables us to understand its possible uses. According to Wittgenstein, the meaning of the words depends on the activity or "language game" it takes place in, and these activities are -often- not governed by strict rules. Therefore, there are no clear-cut ways of determining the semantics of a specific sentence outside the contexts in which it may appear. Thus, the English-speaking linguist in Quine's thought experiment (described above) could assume that the native speaker is actually referring to a rabbit, by also -tacitly- assuming that the native speaker's linguistic community typically has similar everyday behaviors as her/his own linguistic community; or at least a conceptual scheme which is not radically dissimilar.

Perhaps the most related application of these -relatively simple- perspectives on language in the field of NLP concerns word, phrase, and sentence embeddings. Syntactical and semantical relationships between words, phrases, and sentences are learned from the multiple degrees of similarity which they exhibit, and specific grammatical or linguistic rules are not taken into account; only the sheer number of real cases of linguistic use, in the form of textual data. Interesting examples of the learned semantic relationships can result from simple operations (addition or removal) of learned -word- vectors, such as: $V_{Paris} - V_{France} + V_{Italy} = V_{Rome}$ [27, 28]. This example shows that embeddings are able to capture the fact that Paris is the capital of France and Rome is the capital of Italy, without ever being trained to learn what a capital is or how a specific city relates to a specific country (not even how an object relates to a subject and a verb).

As regards translation from one language to another, simple cases such as the translation of the sentence "Water!", which we mentioned above, may be unproblematic. Its translation in Greek, i.e., "Νερό!", is probably correct in all practical settings, i.e., irrespective of the "language games" in which it may be used; although even this simple utterance may be used as a language-specific pun and lose its initial intention after its translation. Especially for longer or specialized sentences, finding the "family resemblances" or the multiple degrees of similarity of the contexts in which certain words of L_{src} are translated to certain words of L_{trg}, is no easy task for an automatic system which is based on texts (or even other modalities) alone and is not directly connected to everyday linguistic activity.

The Transformer NMT model, that will be described in section 2.1 and remains the state-of-the-art in MT at the time of writing, is constituted by two sub-models:

(a) the encoder, which extracts contextual features from a sentence in L_{src}, and,

(b) the decoder, which uses these features to produce a translated sentence in L_{trg}.

Both employ mechanisms to dynamically "pay attention" to potentially distant parts of a sentence when translating. Thus, through the lenses of Willard Van Orman Quine and Ludwig Wittgenstein, the Transformer model could be interpreted as a means of better capturing networks of contextual relationships that emerge in linguistic activities of each

language and, simultaneously, mapping those relationship networks from one language to the other. Put more simply, we can think of the Transformer as a model which is able to map the semantics of a sentence in L_{src} to a sentence in L_{trg}, by better taking context into account via a learned proxy function.

The abovementioned philosophical considerations, in relation with the task of adapting Transformer models to the domain of COVID-19, motivated us to: (a) collect a large number of parallel sentences related to COVID-19 from different contexts (i.e. websites of national authorities, public health agencies, broadcast portals, etc.) and not base our approach solely on data augmentation techniques leveraging monolingual corpora (see Section 2.2), and (b) experiment on the use of "covidity" for selecting which data will be used for adapting our systems, i.e., determine which parallel data are connected with COVID-19 through a specific terminology (see Section 2.3.2). Thus, we hypothesize that, for example, in the pursuit of adapting a NMT system for the translation of scientific publications on COVID-19 (and in the absence of such parallel data), exposing the model to translations in an academic context could be as important as exposing it to translations which are specifically related to COVID-19, even if they originate from non-technical texts.

2. NEURAL MACHINE TRANSLATION AND DOMAIN ADAPTATION

In this chapter, we will provide theoretical background for the task of Neural Machine Translation with the use of the Transformer architecture (see Section 2.1) and give a brief review of Domain Adaptation (see Section 2.2), while also connecting our methodology with Crisis MT and the task of adapting an existing NMT system specifically in the context of the COVID-19 pandemic (see Section 2.3).

2.1 Neural Machine Translation with Encoder-Decoder Networks

Machine Translation is the task of automatically translating texts between two natural languages. Neural Machine Translation has achieved remarkable results in translation quality in recent years. NMT models are given an input sentence x from a source language L_{src} and generate an output sentence y in target language L_{trg}. Sentences x and y are sequences of tokens: a token may take various forms such as a word ("geology"), a character ("g"), or a subword unit ("geo") and it belongs to a predefined vocabulary V along with other tokens.

$$x = \{x_1, x_2, \dots, x_M\}, x_m \in V_{src} \quad (1)$$

$$y = \{y_1, y_2, \dots, y_N\}, y_n \in V_{trg} \quad (2)$$

The Transformer model architecture that we use throughout this thesis follows the encoder-decoder [29] paradigm: sequence x is encoded to a representation which is then decoded to sequence y. In essence, an encoder-decoder model learns the conditional distribution of a variable-length sequence y, given another variable-length sequence x, i.e., $p\big((y_1, y_2, \dots, y_N | x_1, x_2, \dots, x_M)\big)$.

Both the encoder and the decoder are typically either recurrent (RNN), convolutional (CNN), or self-attention neural networks and process each input sequence through multiple stacked layers, with each layer carrying out its operations on the output of the previous layer [30].

2.1.1 Scaled Dot-product and Multi-head Attention

NMT models which utilize attention mechanisms have been established as the state-of-the-art in NMT after their introduction in [31], [32], and [33], due to the great improvements that they achieved, as well as their parallelizability. Their distinctive characteristic is the mechanism of self-attention and, in particular, multi-headed attention, which enables Transformers to dynamically learn and use contextual relationships between -potentially distant- tokens in input sequences and, thus generate more accurate output sequences. In section 1.4, we gave a possible explanation as to why this may be a crucial factor for their success.

An attention function typically maps a query and a set of key-value pairs to an output. The scaled dot-product attention introduced in [33], returns a value vector V weighted by a similarity (or compatibility) function between query vector Q and key vector K (where d_k is the dimension of the keys) according to Equation 3:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Multi-head attention uses h different attention “heads” to linearly project the attention on each input h times, each time using a different learned projection. The outputs are concatenated into a multi-headed attention embedding which represents joint attention on various positions of the sequence. The scaled dot-product attention and the multi-head attention are depicted in Figure 1 which originates from [33], i.e., the original paper which introduced them.

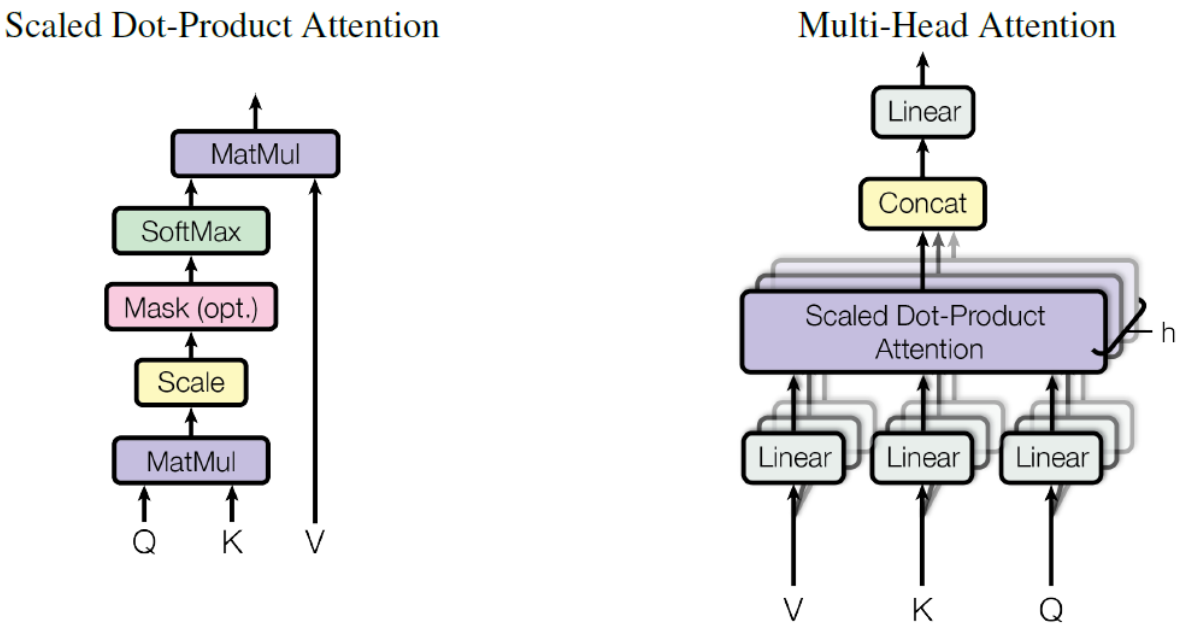

Figure 1: On the left, we can see the scaled-dot product attention, while on the right, we can see the multi-head attention taking place for h heads in parallel.

2.1.2 Transformer Architecture

After providing an overview of encoder-decoder networks and multi-head attention, a general outline of the Transformer model architecture (see Figure 2) can now be given:

- Transformers are comprised by two sub-networks, the encoder and the decoder, which contain identical stacked layers. The final layer of the encoder is given as input to the decoder, while the final layer of the decoder produces the output.
- Each layer of both the encoder and the decoder two sub-layers, a multi-head self-attention mechanism and a fully connected feed-forward network. The decoder includes an additional sub-layer which performs multi-head attention over the output of the encoder sub-network.
- Around each sub-layer, there are residual networks giving each layer access to the original sub-network input [34], and they are followed by layer normalization [35].
- The input and output tokens are converted into vectors using learned token embeddings. Information about the order of the sequence is given to the model through positional encodings (pre-defined or learned functions) which are directly added to the token embeddings [30].

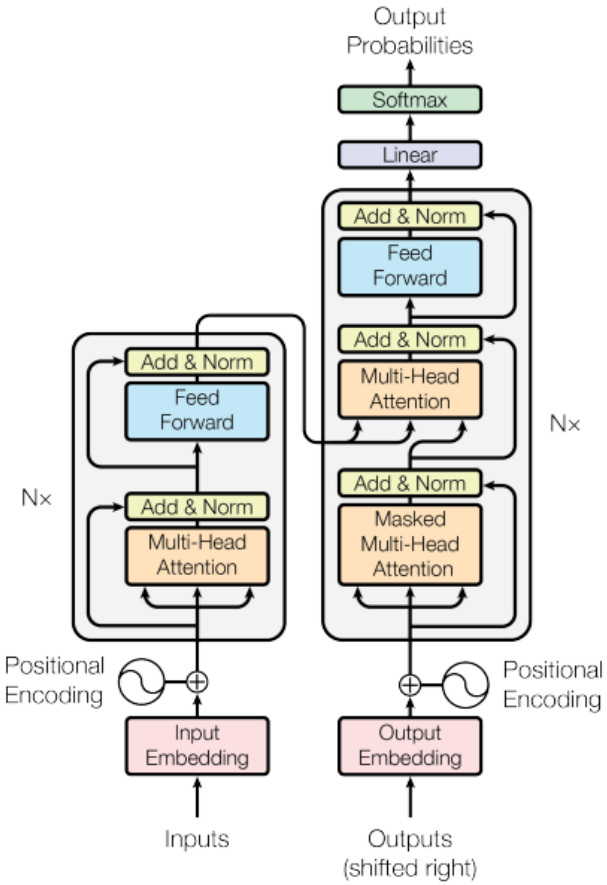

Figure 2: The architecture of the Transformer model

2.1.3 Pre-processing Natural Language for NMT

Up until this point, we have outlined the task of NMT, as well as the Transformer architecture and its various components. As mentioned earlier, the inputs and outputs of the Transformer model are sentences x and y, in the form of sequence of tokens (see Equations 1 and 2) and tokens may take various forms. These forms, however, are limited to specific pre-defined vocabularies. It is apparent that the ability of NMT systems to generalize on new, unseen sentences depends on their capability of open-word translation, i.e., translating rare words and representing the highest possible number of tokens they encounter. Creating a very large vocabulary for a specific language would incur higher computing requirements and out-of-vocabulary words would still be inevitable [30]. Additionally, increasing the size of the vocabulary means that most of its individual instances would be encountered rarely during training of a NMT system and, thus, the system would not learn efficient embedding representations.

Subword units were proposed in [36], towards addressing open-vocabulary translation, and their application has led to consistent improvements in generalization capability for rare and unknown words. This holds especially true for named entities (locations, celebrities, politicians, companies, etc.), words with a common origin (e.g. scientific and medical terms with Greek and Latin prefixes and suffixes), or morphologically complex words such as those found in the German language (e.g. klangkuenstler, klangkarussell, weltschmerz, sternenkinder, etc.) [36].

Throughout the experiments in this thesis, we use vocabularies obtained using the Byte Pair Encoding (BPE) algorithm [36, 37], where each word is split into the most common character sequences of the training dataset. Scripts of the SubwordNMT toolkit [36] are employed for this task, while scripts for tokenization, punctuation normalization, and removal of non-printable characters from the Moses toolkit [38] are also used.

Furthermore, the number of possible subword combinations is reduced dramatically if lowercasing or a truecasing model (which determines the correct capitalization of unknown words) is applied on the original sentences. While the casing method is usually not explicitly stated in many works and experiments, it is common practice to use the corresponding scripts in Moses [38], which train a truecasing model and apply it for case conversion.

Another casing conversion method, which has been shown to be optimal and is gaining more and more popularity, is that of inline casing [39, 40, 41]. Inline casing uses specific tags to denote uppercase (<UC>), title case (<TC>) or mixed case (<MC>) tokens. These tags are placed before or after the tokens which they provide information for and are applied to both sides of a sentence pair. Thus, the NMT system also learns to convert cases as an additional task during training.

2.2 Domain Adaptation

Domain Adaptation in the context of MT should be understood as any scheme aiming to improve the translation quality of an existing system for a certain topic (news, sports, etc.) or genre (technical manual, academic text, etc.) [3]**.** For example, the Biomedical Translation shared tasks which have been organized by the Conference on Machine Translation (WMT) since 2016, aim at evaluating the effectiveness of MT systems for texts in the biomedical domain [14]. While we could think of biomedical texts as a relatively well-defined topic, we could also differentiate between translating biomedical scientific abstracts, clinical case reports, or medical patents, as tasks related to specific genres; these texts serve different functions and may be radically dissimilar in terms of style and syntax.

From all the above, it is evident that the domain of a corpus can -usually- be inferred by its origin, as well as that the concept of “domain” is relative to the particular use-case we aim to solve. For example, SciPar [6], described in section 3.2.3, is a parallel resource collected from parallel abstracts found in academic repositories. If we were interested in adapting a MT system to better translate scientific texts or research summaries (genre), then the corpus could be considered as in-domain. However, if we were interested in adapting a system for a particular subject, such as physics (topic), then only a particular subset of SciPar could be considered as in-domain. Moreover, a NMT system aimed at translating medical news (combination of genre and topic) would probably benefit from medical scientific texts, in addition to data originating from news articles (see Section 1.4).

There are many approaches and methods to adapt a NMT system to a particular domain or even multiple domains. Naturally, the context of each use-case plays a large role in deciding which techniques are useful, as, for example, there may arise circumstances in which we are interested in translating a new domain (e.g., COVID-19) for medium- or low-resource language pairs (e.g., English-Albanian or Greek-Ukrainian). A more in-depth survey of the different types of domain adaptation techniques (e.g., data-centric, architecture-centric, etc.) can be found in [3] and [30]. In the subsections that follow, we present the techniques that are most relevant for our work.

2.2.1 Fine-tuning and Mixed Fine-tuning

Fine-tuning is the conventional and simplest way for domain adaptation, often acting as the baseline for other methods. Fine-tuning consists of continuing the training of an existing NMT system with in-domain data until convergence. There also exist several variations of fine-tuning, such as mixed fine-tuning, which uses a mix of (usually oversampled) in-domain and out-of-domain data [42]. The problems associated with fine-tuning relate to the usually small size of the in-domain corpora and the tendency of fine-tuned models to overfit on the fine-tuning datasets. The latter may lead to the phenomenon of "catastrophic forgetting", i.e., significant translation quality degradation of the model for out-of-domain data [3].

2.2.2 Back-translation and Tagged Back-translation

Back-translation has been proven as an effective data augmentation technique for improving the performance of NMT systems [43, 44] and is particularly useful for domain adaptation and low-resource settings [39]. Methodologically, it consists in generating additional synthetic parallel training data for a L1→L2 MT system (e.g. English→Greek), by translating monolingual sentences in language L2 with a reverse L2→L1 model (e.g. Greek→English). A robust variant of back-translation is tagged back-translation [45] which is easily implemented by inserting a <BT> tag in the beginning of each source sentence which has been synthetically generated. In essence, the model learns to distinguish between the characteristics of originally parallel and synthetic sentence pairs and, when given input sentences without the <BT> tag, it generates translations with statistical characteristics closer to the originally parallel data. In section 3.2.4 we describe how we collected monolingual corpora that were used for back-translation.

2.2.3 Selection of Additional In-domain Data

When in-domain data are scarce, one may use several techniques to extract relevant sentence pairs from other out-of-domain corpora. A category of such techniques is based on content/lexical overlap with in-domain data, while other approaches are based on domain classifiers or similarity scores between sentence embeddings [3]. In this thesis, we used a simple technique primarily to categorize whole datasets according to a pre-defined vocabulary related to the domain of interest (see Section 2.3.2), while we also experimented with extracting in-domain sentence pairs from out-of-domain corpora (see Section 5.3).

2.3 Crisis Machine Translation and COVID-19

As mentioned in section 1.3, Crisis MT can prove beneficial in many crises scenarios, even if they exhibit different characteristics. It is an open question, however, as to what Crisis MT amounts to, as it could be viewed just as a special case of rapid domain adaptation on pre-existing MT systems. Nevertheless, Crisis MT should be better interpreted as a branch of Crisis NLP [46], i.e., the effort of conducting research and developing models, techniques, and datasets specifically tailored for humanitarian aid responses or for improved informational flows towards the most vulnerable parts of the population. Therefore, Crisis MT is by definition human-centric, may be relevant under specific time constraints, and its ultimate evaluation concerns the degree of aid it can provide. In what follows, we describe what we believe to be the essential elements of rapid response NMT systems, as well as the framework we use to conduct research on the case of the COVID-19 pandemic.

2.3.1 Elements of Crisis MT Systems

When building MT systems for crises scenarios, there arise several practical issues that must be addressed before any actual model training. First, the content which is more useful to be translated for the crisis at hand must be identified, as it is possible that this content is not very narrowly related to -what is initially thought to be- a highly specific domain [4]. For example, during the COVID-19 pandemic, governmental information focused largely on hygiene guidelines, social and medical safety precautions, as well as on raising awareness about the safety and efficacy of vaccines. Relevant monolingual (e.g., guidelines, sample medical interactions, frequently asked questions) and parallel data (e.g., biomedical parallel corpora, relevant translation memories, bilingual terminologies) already existed before the emergence of COVID-19. Familiarization with the needs that we aim to address may save valuable time and allows us to select useful data from existing resources, as highly-specific data may be rather scarce or non-existent.

Another important element of a Crisis MT system concerns the infrastructure which could support the actual stakeholders. For example, during the Haiti earthquake in 2010, an SMS messaging infrastructure integrated into a crowd-sourced translation infrastructure was used and proved to be crucial [4]. Practical issues about the infrastructure which could address the needs that were constantly emerging during the COVID-19 pandemic is beyond the scope of this thesis. However, we believe that it is imperative to construct such an in infrastructure in a way that available translation services (e.g., existing commercial services or publicly available models) are integrated and a newly constructed system can promptly take their place when we decide that it is ready for deployment.

In this thesis, we mainly focus on the basic steps that we believe to be essential in the construction of a sufficiently effective MT system which can be integrated in a rapid crisis response infrastructure:

- Familiarization with the target domain and acquisition of existing data which may be relevant.
- Selection and/or acquisition of domain-specific monolingual and parallel corpora.
- Creation of domain-specific developer, test, and generalization sets.
- Evaluation of a -sufficiently good- baseline system on the test set.
- Outline of domain-adaptation strategies (e.g., fine-tuning order, mixed fine-tuning, use of synthetic in-domain data) and possible goals before model deployment such as translation quality; in real settings, these may also relate to model size and speed, temporal constraints, etc.
- Experimentation and comparison of different strategies using in-domain and out-of-domain (to assess over-fitting) developer and test sets.
- Further assessment, improvement, and/or comparison with other translation services using generalization sets.

In real settings, this process is continuous and, of course, specific components can or need to be modified according to the duration, needs, and character of the crisis in hand.

Based on the experience of previous crises as well as forecasts of possible types of future crises, NMT systems adapted to domains and domain-genre combinations for specific language pairs could be constructed beforehand with the use of close-to-domain corpora. A pre-existing collection of such materials, along with the necessary infrastructure, could prove invaluable in rapidly responding to emerging translation needs.

2.3.2 Data Selection for the COVID-19 Domain

In this thesis, we are particularly interested in the domain of “COVID-19”, a topic which emerged in 2019 due to the pandemic, but which remains relevant up to the time of writing this thesis. Since adapting a MT system to COVID-19 is also a Crisis MT scenario, we are interested in conducting experiments which could give insights on how to best respond to future crises with similar characteristics. Thus, in this section, we describe the method that we used to categorize datasets according to their closeness to the domain of COVID-19, as we consider it a crisis in which relevant and useful data pre-existed its emergence.

There are terms which have been commonly used in COVID-19 news (e.g. “contact tracing”, “reproduction rate”, “herd immunity”, etc.) which may be virtually non-existent in parallel corpora of news articles before 2019. In contrast, such terms could probably be found in older corpora originating from biomedical texts. Our goal is to determine the corpora which could prove more useful in adapting MT systems for pandemic-related texts; a form of data selection.

Thus, we introduce a simple heuristic, “covidity”, to determine how closely related a dataset is to the COVID-19 domain. We distinguish “covidity” into two types: (a) *strict “covidity”* which looks for very specific terms: “covid”, “pandemic”, “coronavirus”, “sars cov”, and (b) *extended “covidity”* which also looks for additional related terms such as “vaccination”, “ventilator”, “epidemic”, “disinfectant”, “inflammatory syndrome”, etc. (see Appendix for the full lists of terms).

We categorize all the datasets that we use for training our NMT systems into 3 domain categories, according to the percentage of their sentence pairs in which a specific term appears:

- **Out-of-domain (OOD):** Corpora which have low percentages of strict “covidity” and extended “covidity” are characterized as *out-of-domain*.
- **Close-to-domain (CTD):** Corpora which have low percentages of strict “covidity”, but a percentage of extended “covidity” higher than 10%, are characterized as *close-to-domain*.
- **In-domain (IND):** Corpora which have a percentage of strict “covidity” which is higher than 10% are characterized as in-domain. It is natural to assume that these corpora will also have high percentages of extended “covidity”.

Although the heuristic what we use is quite simplistic, it is a quick way to capture temporal characteristics of the datasets, as well as their closeness to the domain in question. The distinction between the temporal characteristics of the datasets that we use bears practical importance to Crisis MT, i.e., the task of rapidly adapting MT systems for emerging domains. Thus, in-domain corpora will almost certainly be created after the emergence of a crisis (such as COVID-19), but also after sufficient information has been amassed or synthetically generated, while close-to-domain and out-of-domain datasets could predate the crisis. In section 3.3.2, we will use this heuristic to categorize the datasets, while in section 5.1, we will further outline our experimentation framework involving domain adaptation and Crisis MT.

3. DATASET ACQUISITION AND FILTERING

In this chapter, we describe the datasets that were used throughout this thesis. We are specifically interested in English and Greek monolingual sentences, as well as English-Greek parallel sentences. Various sources were utilized: we gathered several existing corpora from the OPUS repository [47], we acquired additional corpora with various methods, and we generated synthetic data through back-translation (see Section 3.2.4). Additionally, we describe the data acquisition methodology with the use of ILSP-FC [48] and LASER [49] in detail, and provide information about data filtering and the categorization of datasets into out-of-domain, in-domain, and close-to-domain.

3.1 OPUS Parallel Corpora

In order to train a NMT system, large amounts of parallel data are needed. There already exist several parallel corpora which have been created for this reason and many of them can be found in the OPUS repository [47]. Since English-Greek is a mid- to high- resource language pair, there exist several open datasets in OPUS which originate from various sources and cover a large range of domains. We obtained the English-Greek sentence pairs from 12 such parallel corpora in OPUS:

- **DGT-TM:**
  The translation memory of the European Commission's Directorate General for Translation. It is a parallel multilingual corpus originating from the European Union's legislative documents in 24 EU languages [50].
- **ECB:**
  Parallel corpus from the website and documentation of the European Central Bank [47].
- **EUBookshop:**
  A multilingual corpus which has been constructed from the EU bookshop, a service and archive of publications from various European institutions [51].
- **Europarl:**
  A parallel corpus which has been extracted from the proceedings of the European Parliament, dating back to 1996 [52].
- **GlobalVoices:**
  A parallel corpus which is constituted from news stories gathered from Global Voices, a website where volunteers have been posting and translating media stories since 2014 [47].
- **SETimes:**
  A parallel corpus of news articles in Balkan languages, extracted from the website of South-East European Times, which published current events in the Balkans from 2002 up to 2015 [53].
- **Wikipedia:**
  Parallel corpus extracted from Wikipedia using web crawling, intermediate translation engines, and extensive filtering [54].
- **OpenSubtitles:**
  A huge and massively multilingual corpus compiled from a large database of movies and TV shows subtitles [55]. It has been updated and re-released multiple times and,

in this thesis, we the 2018 version which was extracted from 3.7 million subtitles in 60 languages [56].

- **TED2020:**
  Another massively multilingual corpus which has been created from the translated subtitles of about 4,000 TED and TED-X talks, available in 100 languages. The transcripts have been created by a global community of volunteers and were crawled in July 2020 [57].
- **WikiMatrix:**
  WikiMatrix is another parallel corpus extracted from Wikipedia, which is larger and covers significantly more languages than others. Its extraction method was based on the use of the LASER toolkit (see Section 3.2.2) by following the global mining approach, i.e., comparing monolingual corpora in pairs and not making use of pre-aligned documents. We use the version with a LASER score threshold of 1.04 [58].
- **CCAligned:**
  The CCAligned corpus was originally created as a dataset of aligned document pairs by matching URLs from 68 snapshots of Common Crawl [59]. In contrast with WikiMatrix, it followed a local mining approach, i.e., the LASER toolkit was utilized to mine parallel sentences from aligned web documents. In this thesis we use a version of CCAligned with a LASER score threshold of 1.08.
- **CCMatrix:**
  The CCMatrix corpus is the largest dataset used in this thesis, as it contains 10.8 billion sentence pairs across 1,197 language pairs in total [60]. It is based on the data released by the Common Crawl project, like CCAligned, and it follows the global mining approach, like WikiMatrix. Similarly with CCAligned, we use a version of CCMatrix with a LASER score threshold of 1.08.

We can see that none of the above corpora is close to the domain of COVID-19, according to their provenance. In section 3.3.2, we will provide relevant statistics which show that this is -mostly- the case indeed.

3.2 Acquisition of Additional Parallel Corpora

In addition to out-of-domain parallel corpora gathered from OPUS, we also acquired additional corpora which are closer to the domain of COVID-19. The acquisition and processing of some of them involved a large part of work, since corpora closer to the domain that we are interested in were rather scarce. Thus, we acquired various parallel corpora for the English-Greek language pair, although it should be noted that the acquisition process that we describe targeted significantly more language pairs in European languages. In the sub-sections that follow, we describe the two most important components of our parallel data acquisition approach, the identified and newly created corpora, as well as the methodology used to generate synthetic parallel sentences related to COVID-19.

3.2.1 Crawling Parallel Documents with ILSP-FC

The ILSP-FC (Institute for Language and Speech Processing - Focused Crawler) toolkit [48, 61] was used extensively for the acquisition of domain-specific corpora from the web. Essentially, ILSP-FC is a pipeline to acquire domain-specific monolingual and bilingual documents from the web. It follows a modular architecture and, thus, each of its modules

can be replaced by another one with the same functionality [61]. Its main components are the following:

- **Page Fetcher:** Crawls a website by concurrently visiting its pages and documents, beginning from a starting seed URL list and proceeding in cycles to visit links according to the crawling strategy dictated by the Link Extractor module.
- **Normalizer:** Detects the text encoding of the web page, converts it to UTF-8 (if needed), parses its structure and extracts its metadata.
- **Cleaner:** Segments texts into paragraphs based on HTML tags, identifies boilerplate components (e.g., headers, disclaimers, etc.), and extracts structural information.
- **Language Identifier:** Detects the language of the document and its paragraphs, as a web page may include text in multiple languages. If a web page is not in one of the targeted languages, its only use is the extraction of more links for the Page Fetcher module.
- **Domain Checker:** Compares the contents of a web page with an optional, user-provided domain definition, calculates the domain-relevance score of each web page, and categorizes it as relevant to the domain if that score is above a pre-defined threshold.
- **Link Extractor:** Examines the links that have been extracted from a web page and ranks them according to how “promising” they are. The list of URLs that will be visited by the Page Fetcher module in the next cycle gives priority to links that are more probable of being candidate translations of web pages fetched in the current cycle, or related to the domain that we are interested in.
- **Exporter:** Exports text and metadata from each stored web document and stores it in XML format. These files are enriched with structural information, i.e., information about the paragraphs of a document, as well as annotations indicating text pieces which are headings, titles, boilerplate, not in any of the targeted languages, etc.
- **De-duplicator:** Compares each document with all others, identifies its -near-duplicate documents and keeps the longest one. The De-duplicator module also takes word frequencies and the percentage of common paragraphs into account.
- **Pair Detector:** Detects parallel documents by exploiting extracted metadata, URL patterns, co-occurrence of the same image filenames (while not considering images that appear frequently), sequences of digits and structural similarity. It is worth mentioning that the Pair Detector module is language agnostic.
- **Sentence Aligner:** Uses an open-source aligner to extract segment pairs from parallel documents and generates a TMX (Translation Memory Exchange) file with the segment pairs of each document.

In Figure 3, which has been acquired from [61], we can see a depiction of the ILSP-FC workflow. ILSP-FC was used to extract parallel and monolingual documents from websites of national authorities and public health agencies [62], European and intergovernmental agencies, and specific news portals. It was also used to extract a Greek monolingual corpus related to COVID-19 from the web, which was used for back-translation (see Section 2.3.2).

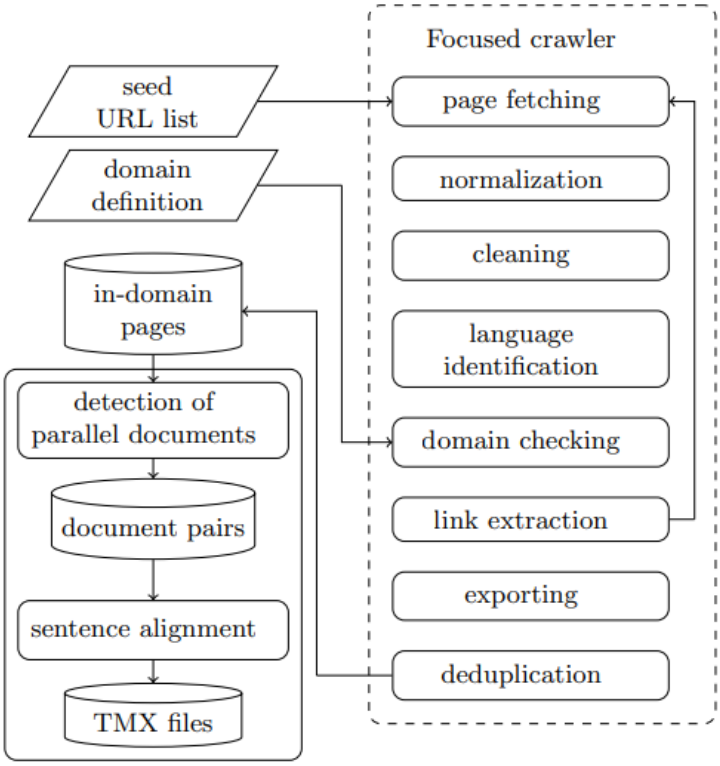

Figure 3: Overview of the ILSP-FC pipeline

3.2.2 Mining Parallel Sentences with LASER

The LASER toolkit [49] was used extensively throughout data acquisition and filtering, as it provides a straightforward method to mine sentence pairs from parallel or even comparable texts, while it also assigns a sentence alignment score to each sentence pair (henceforth referred to as “LASER score”). The LASER score is a corpus-dependent metric which assesses the semantic similarity between sentences in different languages [63].

Mining sentence pairs from parallel or comparable documents was traditionally based on heavily engineered and language-dependent systems. In contrast, LASER uses a pre-trained multilingual sentence encoder which maps sentences to fixed-length vectors in a common space and, at the time of writing, supports over 200 languages. In Figure 4, acquired from [64], we can see a qualitative example of the approach: When multilingual sentences (in English, French and German) which have similar meanings are embedded in a common space (depicted as three-dimensional), their vector representations are in close proximity and, therefore, have high cosine similarities.

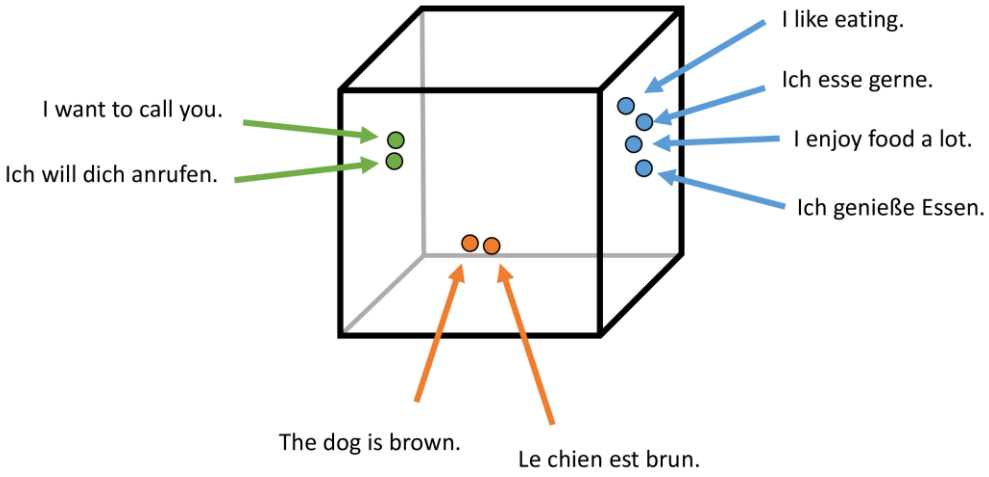

Figure 4: Illustration of embedded multilingual sentences in a common space

For each source sentence x, the k-nearest neighbor (we use the default configuration: k=4) target sentences $NN_k(x)$ are determined in terms of the cosine similarity between the source vector x and the target vectors. Likewise, the k-nearest neighbors $NN_k(y)$ are computed for target sentence y. Afterwards, the LASER score is computed so as to select the best candidate sentence pairs. The LASER score (see Equation 4) is based on a margin function between the cosine similarity of each candidate sentence pair and that of its respective k-nearest neighbors. We follow the authors [49] in using the ratio criterion, i.e., the ratio between the cosine similarity of each candidate sentence pair and the average cosine similarity of its 4 nearest neighbors.

$$score(x,y) = \frac{cos(x,y)}{\sum_{z \in NN_k(x)} \frac{cos(x,z)}{2k} + \sum_{z \in NN_k(y)} \frac{cos(y,z)}{2k}} \quad (4)$$

In order to mine sentence pairs from a web page, we split the parallel documents extracted by the ILSP-FC toolkit (see Section 3.2.1) into sentences, and then apply the margin-based scoring method for each document pair, as well as between the whole monolingual corpora (concatenated sentences of all language-specific documents) extracted from each web page. Thus, our mining method combines the local and global approaches (see Section 3.1). By experientially setting thresholds of the LASER score for each mined corpus, along with applying various filtering methods (see Section 3.3.1), we reduce the noise of the data used for training our NMT systems.

3.2.3 Close-to-domain and In-domain Corpora

As mentioned in section 3.2, in addition to out-of-domain parallel corpora gathered from OPUS, we also acquired additional corpora which are closer to the domain of COVID-19. All of the four corpora that are described in what follows, have been constructed by the ILSP and this thesis has contributed to the construction of the latter three.

- **EMEA:**
  The European Medicines Agency (EMEA) website [65] contains many PDF documents in the official EU languages and thus, constitutes a promising resource of high-quality multilingual data in the biomedical domain. In this thesis, we use an improved version of the EMEA parallel corpus which has been constructed by the ILSP and is available on the ELRC-SHARE repository [66].

- **CovidWebsites:**
  This multilingual corpus was acquired using the methods described in sections 3.2.1 and 3.2.2 and contains COVID-19 related information in most European languages, gathered from websites of national authorities and public health agencies, as well as intergovernmental (e.g., EU and UN) sources [9, 10, 62].
- **MediSys:**
  MediSys is a parallel resource which has been constructed by exploiting the COVID-19 related datasets of metadata created with the EMM - MediSys (Europe Media Monitor – Medical Information System) processing chain of news articles [8]. In total, it contains 11.2 million sentence pairs in 26 EN-X language pairs which have been mined from comparable monolingual corpora using the methods described in section 3.2.2 [7]. Although the original version of MediSys has been filtered extensively, in this thesis we use the raw version of the EN-EL parallel corpus since we apply different filtering methods (see Section 3.3.1).
- **SciPar:**
  SciPar is a collection of parallel corpora which has been constructed from bachelor theses, master theses, and doctoral dissertations available on institutional repositories, digital libraries of universities, and national archives [6]. It comprises 9.17 million sentence pairs in 31 language pairs, which have been mined with the methods described in section 3.2.2, and is publicly available via the ELRC-SHARE repository [66]. Although it includes data from multiple academic fields and sub-disciplines (such as medicine, informatics, engineering, philosophy, etc.), it addresses the relatively under-resourced genre of scientific research and, thus, it may prove useful in improving the translation of, e.g., COVID-19 related publications (see Sections 1.4 and 2.2).

It is worth mentioning that the CovidWebsites and MediSys corpora, that we described above, have been used at the COVID-19 MLIA-Eval [9, 10, 67], an initiative focused on facilitating the development of language resources and tools for COVID-19 towards improving multilingual information access. COVID-19 MLIA Eval includes three shared tasks open to participants: (a) information extraction from medical texts, (b) multilingual semantic search, and (c) machine translation of COVID-19 related texts. Monolingual corpora from the aforementioned corpora were used for the multilingual semantic search task, while the parallel corpora were used for the machine translation task. More information on the machine translation task can be found in [9] and [10].

3.2.4 Generation of Synthetic Parallel Corpora

Parallel corpora related to COVID-19 are not only rather scarce, but also require a significant amount of time to be compiled. When we are faced with a crisis and we want to adapt a NMT system for related information, we may not be able to construct adequate domain-specific parallel corpora promptly; time is of the essence. Back-translation (see Section 2.2.2) is a simple, fast, and robust technique to use for domain adaptation in such cases. It consists of leveraging monolingual corpora (which can be obtained faster) to generate synthetic parallel corpora using a reverse NMT system.

In this thesis, we use the ILSP-FC toolkit (see Section 3.2.1) to gather monolingual Greek sentences related to COVID-19. Towards this end, we utilize the Domain Checker module of the ILSP-FC which crawls web pages according to a relevance score assorted with the weighted occurrence frequency of health and COVID-19 related terms [61]. In total, **658,464** Greek sentences were obtained and were translated into English with the Google Cloud

Translation AI API [68] (General NMT model v3). In the sections that follow, we call this synthetic parallel corpus "**BackTranslatedSmall**" for convenience.

As regards the additional experiments towards building an end-to-end NMT system (see Section 6.1), we gather **27,550,834** Greek sentences originating from various sources, such as the monolingual Greek corpus of MediSys [7], further crawling with the ILSP-FC toolkit (see Section 3.2.2), other resources available in the CLARIN:EL infrastructure [69], etc. Thus, a larger synthetic parallel corpus is generated with the utilization of the Hugging Face library [70] and, in particular, with the EL→EN SSE-TUC model [71]. This generated synthetic parallel corpus is hereafter referred to as "**BackTranslatedLarge**".

3.3 Dataset Filtering and Domain Categorization

In this section, we describe the filtering methods that were applied on the datasets used in this thesis and the categorization of each one into in-domain, close-to-domain, or out-of-domain, according to section 2.3.2. Additionally, we report the filtering and "covidity" statistics for each dataset.

3.3.1 Filtering Methods for Parallel Corpora

Since the parallel corpora that we use originate from various sources (most originate from OPUS, some have been constructed by us, while others have been synthetically generated) and have been mined and filtered with different methods, we believe that it is quite useful to apply a common filtering pipeline. Filtering has been shown to improve the performance of NMT systems, even if the applied filters are simple [72], since it reduces the noise in the parallel corpora. Although there have been significant developments in filtering methods based on neural networks, we decided to apply a series of simple rule-based filters which are more easily interpretable. The following pipeline has been applied on all the parallel data, including the synthetic data:

- Sentence pairs in which the source and target sides are identical or either side is empty are removed [72, 73].
- Duplicate sentence pairs are removed, based on either the source or the target side; i.e., no English or Greek sentence appears more than once in the training set. Sentences were normalized (i.e., were lowercased and had digits and punctuation removed) before the deduplication process [5].
- Sentence pairs in which either side consists of more than 50% non-alphabetic characters are removed [74].
- Sentence pairs in which the length ratio (between the two sentences) in terms of digit characters is over 2:1 (or below 1:2) are removed [5].
- Sentence pairs with either sentence in a language different than expected are removed. Language identification is performed using fastText [75].
- Sentence pairs in which either the source or the target normalized sentence contains less than 3 or more than 120 tokens are removed. This configuration was initially unproblematic in the domain adaptation experiments (see Section 5.1). However, in experiments involving the construction of a complete end-to-end NMT system (see Section 6.1), in which our end-goal is a generalizable model, we decided to change the configurations and only remove sentence pairs with less than 1 or more than 250

tokens. Normalized sentences with less than 1 token are solely constituted by digits and punctuation.

- Sentence pairs in which the token ratio between the longest and shortest sentence is higher than 2, after normalization, are removed [5].
- The repeating token filter is applied in order to remove low-quality machine translated content which in many cases tends to produce repeating sequences of the same token, if the system is unable to translate a part of the sentence [74].
- Sentence pairs in which either normalized side contains letters not in the range of Unicode character sets relevant to Greek and Latin, are removed [73]. This aims to remove characters which are not in the languages of interest or other kinds of non-standard symbols. Experiments with this filtering method, which was only applied in the end-to-end NMT system (see Section 6.1), showed that the subword vocabularies (see Section 2.1.3) needed to train our models can be reduced in size, thus reducing the number of model parameters.

In Table 1, we report on the starting (“raw”) sizes of the EN-EL parallel corpora, the total sentences pairs which remained after the application of the filtering pipeline that we described above, as well as the percentage of the initial sentence pairs which remained after filtering.

Table 1: Filtering statistics of the EN-EL parallel corpora

EN-EL Corpus	Raw	Filtered	Percentage
DGT-TM	5,099,790	1,530,979	30.02%
ECB	102,986	31,155	30.25%
EUBookshop	4,022,952	1,474,255	36.65%
Europarl	1,292,180	918,611	71.09%
GlobalVoices	120,421	76,832	63.80%
SETimes	227,168	180,551	79.48%
Wikipedia	104,076	53,804	51.70%
OpenSubtitles	40,492,942	13,253,102	32.73%
TED2020	269,407	171,365	63.61%
WikiMatrix	620,802	452,252	72.85%
CCAligned	3,579,301*	1,443,011	40.32%*
CCMatrix	19,960,688*	10,347,475	51.84%*
EMEA	781,197	487,649	62.36%
CovidWebsites	153,392	65,470	42.68%
MediSys	529,518	339,651	64.14%
SciPar	742,986	553,170	72.88%
BackTranslatedSmall	658,464	471,495	71.61%
BackTranslatedLarge	27,550,834	20,636,616	74.90%

***:** The raw sizes of CCAligned and CCMatrix that we list in Table 1, concern the versions with a LASER score threshold of 1.08 (see Section 3.1). Accordingly, the percentage of the initial size remaining concerns these sizes.

3.3.2 Covidity Statistics and Categorization

In section 2.3.2, we described the measures of strict “covidity” and extended “covidity” that we use in order to categorize each parallel corpus into three domain categories: out-of-domain (OOD), close-to-domain (CTD) and in-domain (IND). We calculated the percentages of these two measures in each dataset by checking if the corresponding pre-defined terms (see Appendix) appear in the English side of each sentence pair. We list these two measures in Table 2 and assign a domain category to it according to the criteria

in section 2.3.2: CTD corpora have >10% extended “covidity”, IND corpora have >10% strict “covidity”, and OOD corpora have <10% of both measures. We can observe that all the corpora from OPUS are categorized as OOD, two parallel datasets are categorized as CTD, while four datasets (including the two synthetic ones) are categorized as IND. This categorization is used in the domain adaptation experiments in section 5.1.

Table 2: Information on covidity and domain category of the EN-EL parallel corpora

EN-EL Corpus	Strict Covidity	Extended Covidity	Domain Category
DGT-TM	0.00%	3.84%	OOD
ECB	0.00%	1.25%	OOD
EUBookshop	0.00%	3.11%	OOD
Europarl	0.02%	3.92%	OOD
GlobalVoices	0.00%	2.91%	OOD
SETimes	0.01%	2.72%	OOD
Wikipedia	0.01%	3.30%	OOD
OpenSubtitles	0.00%	1.03%	OOD
TED2020	0.06%	4.15%	OOD
WikiMatrix	0.01%	3.09%	OOD
CCAligned	0.02%	1.90%	OOD
CCMatrix	0.02%	4.98%	OOD
EMEA	0.12%	37.67%	CTD
CovidWebsites	10.88%	29.55%	IND
MediSys	13.56%	32.71%	IND
SciPar	0.09%	14.39%	CTD
BackTranslatedSmall	18.89%	60.41%	IND
BackTranslatedLarge	33.00%	69.59%	IND

4. EVALUATION FRAMEWORK

In this chapter, we describe the evaluation framework that was used in the domain adaptation experiments in chapter 5, as well as the comparison of end-to-end NMT systems in chapter 6. In what follows, we give an overview of the metrics and test sets used throughout this thesis.

4.1 Automatic Evaluation Metrics

As we saw in section 1.4, translation is an inherently vague and indeterminate task and, therefore, human judgements (from multiple judges) are needed to thoroughly evaluate a MT system. However, human evaluations are expensive and time-intensive, while MT systems need to be tested and compared on the fly. In order to address this problem, the MT community has proposed countless metrics which aim to better correlate with human judgments. In this section, we will describe three metrics for MT evaluation: BLEU, chrF2++, and COMET.

4.1.1 BLEU and SacreBLEU

BLEU has, beyond any doubt, been the most crucial metric in MT over the last two decades, since its introduction in [76]. In a meta-study of 769 MT research papers published between 2010 and 2020, it was found that 98.8% of the studied papers reported BLEU scores of MT models, even though 108 new metrics which better correlate with human judgments have been proposed [77].

BLEU is a language-independent metric which was based on a quite simple idea: “The closer a machine translation is to a professional translation, the better it is” [76]. For its calculation, it requires a sentence in L_{src}, a candidate translation in L_{trg} generated by a MT system, and a reference translation in L_{trg}. It compares n-grams between the candidate and the reference translation in a position-independent fashion, and counts the matches. In other words, a candidate translation gets a higher BLEU score if it shares many words and phrases with a “good” reference translation.

The variant of BLEU that we use is the detokenized BLEU with a maximum n-gram order of 4, and is implemented in the SacreBLEU toolkit [78], which provides scripts for computing BLEU and chrF2++ scores in a reproducible fashion. For the experiments in this thesis, the BLEU short version signature from SacreBLEU is:

nrefs:1|bs:1000|seed:12345|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0

4.1.2 chrF and chrF2++

The chrF2++ metric that we use in this thesis originates from the chrF or character n-gram F-score metric, introduced in [79], and has been described as the best performing string-based metric [80]. Like BLEU, it is a language-independent metric which has shown good correlations with human judgments. However, unlike BLEU, its initial version (chrF), uses only character n-grams (which also means that it did not depend on the tokenization methods). The version that we use (chrF2++) has a β parameter of 2, i.e., recall has two times more importance than precision [81] and also utilizes word bigrams [82], in addition to character 6-grams. The chrF short version signature from SacreBLEU is:

nrefs:1|bs:1000|seed:12345|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.0.0

4.1.3 COMET

COMET is a neural framework for training multilingual MT evaluation models which serve as pre-trained automatic metrics [83]. At the time of writing, it is the state-of-the-art in MT systems evaluation, as it exhibits the best correlation with human judgments [80]. In this thesis, we use the default model ("wmt20-comet-da") which is a reference-based regression model built on top of XLM-RoBERTa [84], a pre-trained cross-lingual language model. In particular, COMET is used in chapter 6 to compare our end-to-end domain-adapted NMT system with other available models and translation services.

4.2 Domain-Specific Test Sets

The choice of the test sets which we use to evaluate and compare our NMT systems is quite important. As we will see, even though test sets for most language pairs are rather scarce, there are several pre-existing test sets for the EN-EL language pair. If we are particularly interested in assessing the translation quality for a specific domain, we may be required to construct a new test set, as is the case for the COVID-19 pandemic.

Additionally, as we discussed in section 2.2.1, adapting a NMT system using fine-tuning may lead to "catastrophic forgetting", i.e., reduced translation quality for out-of-domain data. Thus, we need at least two test sets: one which is in the domain of COVID-19 and a general one. However, pre-existing general test sets may contain short or generic sentence pairs or may actually be mixtures of specific domains. Therefore, we believe that several test sets (pre-existing and new custom ones) are required so as to effectively analyze the effects of the different techniques used in experiments. We use the following EN-EL sets:

- A **General** test set created by sampling 750 sentence pairs from each of the 12 out-of-domain parallel corpora from OPUS (see Section 3.1) after the application of the filtering pipeline (see Section 3.3.1). More specifically, we sampled 9,000 sentence pairs in total to create (a) a developer set, (b) a test set, and (c) a generalization set, each comprising 3,000 parallel sentences.
- A **COVID-19** related test set (referred to as "COVID-19"), created in a similar fashion as the General set described above, by sampling 2,250 sentence pairs from the 4 filtered close-to-domain and in-domain parallel corpora mentioned in section 3.2.3. In particular, we constructed 3 sets (developer, test, and generalization) with 3,000 parallel sentences each, by ensuring that each sentence contains at least one term from the ones used to calculate extended "covidity" (see Section 2.3.2); in other words, the COVID-19 developer, test, and generalization sets all have 100% extended "covidity".
- The **Tatoeba** test set which contains 10,899 sentence pairs and has been created for the Tatoeba Translation Challenge [85], using crowd-sourced translations. It contains generic and short sentences, some of which are very similar and differ only in terms of grammatical form.
- The **XNLI** (Cross-lingual Natural Language Inference Corpus) test set is constituted by 10,000 sentences from texts and transcripts which have been translated into several language by professional translators for use in cross-lingual language understanding applications [86]. In this thesis, we use the unlabeled English-Greek sentence pairs.

The four test sets listed above, as well as in Table 3, amount to a total of 26,899 sentence pairs and are quite sufficient for the domain adaptation experiments in chapter 5, since we are mostly interested in assessing translation quality for COVID-19 related texts while also

monitoring the emergence of “catastrophic forgetting” on out-of-domain texts for each domain adaptation strategy (see Section 2.2.1).

Nevertheless, as we will see in section 6.1, we also attempt to construct an end-to-end NMT system which is adapted for COVID-19 related data and compare it with other systems. Meaningful comparison of NMT systems, however, would require testing on additional domains and styles of writing, as each system may exhibit its own strengths and weaknesses. Thus, in these experiments, we use four additional test sets, two of which have been constructed as part of this thesis:

- The **TED Talks** test set contains 4,431 sentence pairs originating from TED talks transcripts, i.e., translated spoken data from presentations in TED conferences [87].
- **ACCURAT** is a balanced test corpus for under-resourced languages which is available on the CLARIN:EL portal [88] and consists of 512 parallel sentences.
- The **SciPar** test set has been generated from the dataset of the same name that we acquired (see Section 3.2.3) and originates from theses’ and dissertations’ abstracts in multiple scientific fields [6]. It contains 2,000 sentence pairs from 2,000 distinct parallel documents (abstracts) and thus, captures a broad range of scientific fields.
- The **Literature** test set is constituted by 2,000 sentence pairs which have been mined from the translated versions of various classic literary works, utilizing the methods described in sections 3.2.2 and 3.2.3. We created this test set specifically for this thesis.

In summary, the evaluation framework for end-to-end MT systems makes use of 3 metrics (BLEU, chrF2++, and COMET) computed for 8 test sets comprising 35,842 parallel sentences in total, as listed in Table 3. These test sets feature various styles of texts, as they originate from websites, books, transcripts of presentations and proceedings, spoken dialogues, academic texts, news articles, official EU texts, etc. Furthermore, they cover a wide range of domains such as scientific research, COVID-19, news, literature, legal, biomedical, commercial, etc. Thus, we believe that they can provide us with meaningful insights on system performance.

Finally, it is worth noting that the General and COVID-19 test sets are used in the evaluation of the domain adaptation experiments in section 5.3, while the corresponding generalization sets are used in the evaluation of the end-to-end MT systems in section 6.2. We use distinct sets in these two cases because the choice of the fine-tuning strategy in section 6.1 is based upon the best-performing strategy which resulted from the experiments in chapter 5. Otherwise, if we used the same test set to both select the best-performing system and evaluate it with other MT systems, the final evaluation of end-to-end translation systems would be more biased towards our approach.

Table 3: Number of sentence pairs in the test sets

EN-EL Test Set	Sentence Pairs
General	3,000
COVID-19	3,000
Tatoeba	10,899
XNLI	10,000
Total used in Chapter 5	**26,899**
TED Talks	4,431
ACCURAT	512
SciPar	2,000
Literature	2,000
Total used in Chapter 6	**35,842**

5. DOMAIN ADAPTATION EXPERIMENTS

In this chapter, we outline the framework of the domain adaptation experiments that we conducted in order to compare different rapid domain adaptation strategies on existing systems in a crisis scenario (COVID-19). Additionally, we describe the training and fine-tuning configurations for the NMT systems, provide an analysis of the experimental results, and discuss the most important steps required in Crisis MT.

5.1 Experimentation Framework

During a crisis scenario, we typically already have a general NMT system for a given language pair or several language pairs. The translation needs arising from a crisis can be interpretated as an emerging domain or combination of domain and genre (e.g., legal texts for refugees), depending on the specific characteristics of the situation. Adapting an existing NMT system, especially when there are changes in translation needs and data availability during a crisis scenario, is not a straightforward task. There exist various practical issues regarding domain adaptation strategies which need to be explored. The fine-tuning order and the effects of mixed fine-tuning, for instance, have been identified as very important issues which require experimentation and comparison [89].

With these in mind, we decided to explore various fine-tuning strategies based on the following three phases of a -simulated- crisis scenario:

- **Phase 0:** The crisis scenario has just been identified. Only general-domain NMT systems are currently available (or can be trained), and they have been trained -mostly- on out-of-domain data.
- **Phase 1:** Several characteristics and needs of the crisis scenario have been -at least loosely- identified. NMT systems which have been already adapted to pre-existing domains or domain-genre combinations similar to that of the crisis scenario could promptly become available. Similarly, close-to-domain corpora or -in general- data have been identified at this phase.
- **Phase 2:** Most of the characteristics of the crisis scenario have been identified and an adequate amount of related data has been generated. During this phase, NMT systems adapted for the specific crisis can become available, and in-domain corpora have been constructed or can be created from available resources.

First, we train a baseline NMT system using the 12 out-of-domain OPUS parallel corpora (see Section 3.1), labelled as “OOD” in Table 2. We hereby referred to this system as “Phase 0”. Afterwards, we continue training it with different fine-tuning strategies corresponding to different domain types of data; for example, strategy S2, “Phase 0 → Phase 1 → Phase 2”, consists of fine-tuning the baseline system first with the two CTD corpora and afterwards with the four IND corpora (see Table 2) and would probably constitute the most temporally realistic scenario. Nevertheless, we also experiment with the reverse strategy, S4, “Phase 0 → Phase 2 → Phase 1”, as well using a single fine-tuning step (with both CTD and IND data) in strategy S5, “Phase 0 → Phase 1+2”. In strategy S7 we train an “oracle” NMT system, i.e., a system directly trained with all data. We list the different fine-tuning strategies, as well as the number of sentence pairs used in the training/fine-tuning process below in Table 4.

Table 4: Description of domain adaptation strategies and number of data used in each one

Strategy	Fine-tuning order	Number of sentence pairs used for training/fine-tuning
S0	Phase 0	28,996,683
S1	Phase 0 → Phase 1	1,025,888
S2	Phase 0 → Phase 1 → Phase 2	820,323
S3	Phase 0 → Phase 2	820,323
S4	Phase 0 → Phase 2 → Phase 1	1,025,888
S5	Phase 0 → Phase 1+2	1,846,211
S6	Phase 0 → Phase 1+2+0*	2,661,571
S7	Phase 0+1+2	30,842,894
S8	Phase 0+1+2 → Phase 1+2	1,846,211

***:** Additional 815,360 sentence pairs with extended "covidity" terms have been selected from OOD/Phase 0 data and are used alongside CTD and IND data for a greater degree of mixed fine-tuning (see Section 2.2.1) in strategy S6.

5.2 Architecture and Configuration

In order to train the NMT systems in the experiments, we train Transformer models [33] with a single RTX 2080 Ti GPU using the Fairseq toolkit [90]. The "big Transformer" architecture [33] is used with 6 encoder layers, 6 decoder layers, 16 attention heads (in both the encoder and the decoder), a dimensionality of 1,024 for the input and output (encoder and decoder embeddings) and a dimensionality of 4,096 for the inner fully-connected network.

Using the Subword NMT [36] toolkit and Fairseq [90], we train a joint BPE tokenizer with 32,000 merge operations and, subsequently, build two dictionaries: The English dictionary contains 17,648 subwords (dimensions of the embedding vectors of the encoder) and the Greek dictionary contains 32,096 subwords (dimensions of the embedding vectors of the decoder). These subwords are used to represent all the inputs and outputs of the NMT system.

We apply dropout with probability 0.3, activation dropout with probability 0.1 and attention dropout with probability 0.1. The Adam optimizer [91] is used with a peak learning rate of 0.0005 after 4,000 warmup steps and then follows inverse square root decay. Each training batch contains a maximum of 1,900 tokens and the parameters of the model are updated every 8 batches.

Checkpoints of the model are saved every 20,000 updates when training the models in one step (i.e., in S0 and S7) and every 2,000 updates when fine-tuning. We use a patience of 5, i.e., the training is halted if the BLEU score calculated on the validation set (which is a concatenation of the General and COVID-19 validation sets described in section 4.2) does not improve for 5 checkpoints. In order to obtain the final model parameters, we employ checkpoint averaging [30, 33, 92], and thus average the parameters of the 8 last saved

checkpoints when training in a single step (S0 and S7) and of the 4 last checkpoints when fine-tuning. Beam search 5 is used for inference.

In Figure 5, we can see the training process of the baseline model S0 in terms of the BLEU score, as well as the range of updates (160,000 up to 320,000) that were used for checkpoint averaging and resulted in the final model parameters.

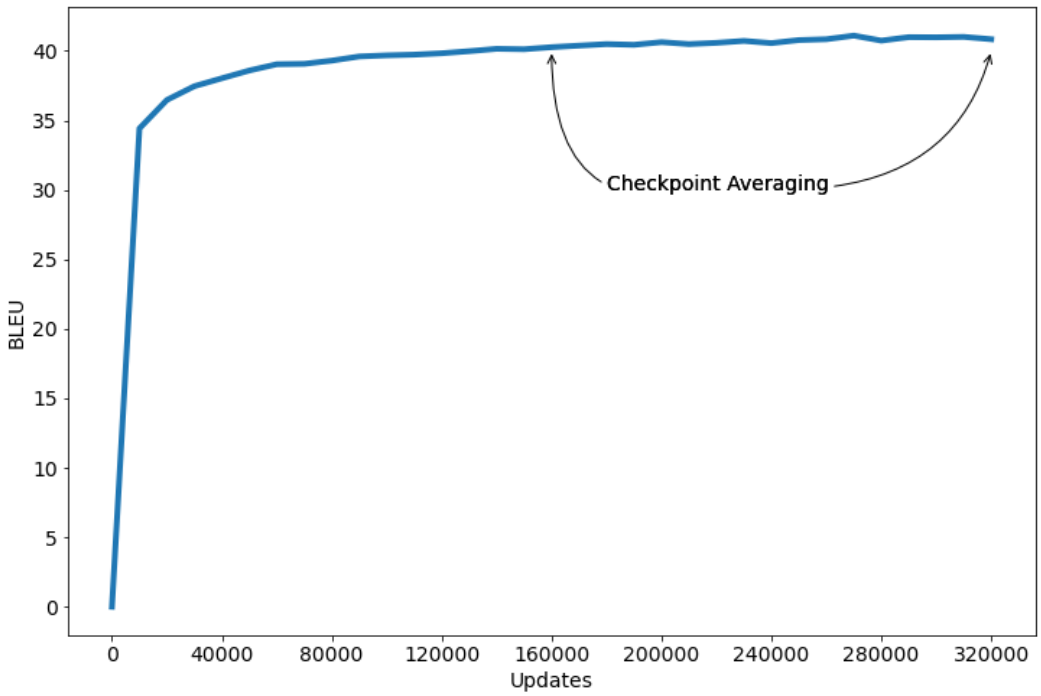

Figure 5: Training progress of the baseline model (S0)

5.3 Results and Analysis

In Table 5, we can see the BLEU and chrF2++ scores computed with SacreBLEU [78] for the four test sets that we use in the domain adaptation experiments (see Section 4.2). The following observations can be made:

- The baseline model S0 outperforms all other models on the Tatoeba and XNLI test sets, with the oracle model S7 having the second best performance on those sets and surpassing the baseline on the General test set; nonetheless, these differences are very small and not statistically significant.
- In strategy S1 (Phase 0 → Phase 1), we fine-tune the baseline model with the CTD data and achieve +2.5 BLEU (and +2 chrF2++) on the COVID-19 test set. Further fine-tuning with the IND data in strategy S2 (Phase 0 → Phase 1 → Phase 2), however, does not lead to a corresponding increase in performance on the COVID-19 test set and positively affects only the results on the Tatoeba test set.
- The less realistic strategies S3 (Phase 0 → Phase 2), in which we fine-tune the baseline model with the IND data first, and S4 (Phase 0 → Phase 2 → Phase 1), in which we continue training with the CTD data afterwards, seem to only worsen the performance on all test sets.
- Overall, the best-performing system on the COVID-19 test set (46.1 BLEU, 65.0 chrF2++) is the one in strategy S8 (Phase 0+1+2 → Phase 1+2), which concerns the fine-tuned oracle system (Oracle-FT) and in which an initial model trained with OOD,

CTD and IND data is fine-tuned with the CTD and IND data. Nevertheless, strategy S8 is only marginally better (+0.4 BLEU, +0.3 chrF2++) than the more realistic (for a crisis scenario) strategy S5 (Phase 0 → Phase 1+2), in which we continue training with a single fine-tuning step with CTD and IND sentence pairs. We can also observe that the effect of "catastrophic forgetting" on the three OOD test sets is greater for S5 as compared for S8 and hypothesize that it is due to the data exposure of the systems; in contrast with S5, the model in S8 has already been trained with the data that it is being fine-tuned with.

- In section 2.2.1, we mentioned the technique of mixed fine-tuning which uses additional OOD data (and usually oversampled IND data) for fine-tuning and results into lesser degradation of general performance. In this thesis, we experimented with a specialized mixed fine-tuning strategy, S6 (Phase 0 → Phase 1+2+0), in which we used 815,360 additional parallel sentences (containing at least one of the terms used to determine extended "covidity") selected from datasets labelled as OOD (see Table 4). This change alone was enough to mitigate the effects of "catastrophic forgetting" to a large degree, since we observe improvements from +0.9 up to +1.7 BLEU (correspondingly, +0.8 to +1.4 for chrF2++) on the three general test sets in comparison to S5 (Phase 0 → Phase 1+2), while the decrease in performance on the COVID-19 test set is not statistically significant (-0.3 BLEU and -0.3 chrF2++).

As regards a realistic domain adaptation strategy for the crisis scenario we investigate, the above observations would lead us to the conclusion that a single step of fine-tuning on a pre-existing NMT system yields the best results and has a similar performance with an oracle model, i.e., a model trained with all relevant data in advance. Indeed, the findings that **(a)** S5 exhibits similar performance on the COVID-19 related test set with S8, and **(b)** S6 exhibits similar performance on the three OOD test sets with S8, show that there is a balance between performance on in-domain and out-of-domain texts which can be better achieved with the implementation of an appropriate fine-tuning order and mixed fine-tuning. Nevertheless, further research is needed on mixed fine-tuning strategies, possibly with oversampled IND data, and domain control using domain tags (i.e., <IND> or <OOD>) at the beginning of each input sentence [93].

Table 5: BLEU and chrF2++ scores on the 4 test sets for each domain adaptation strategy

	General		Tatoeba		XNLI		COVID-19	
Strategy	**BLEU**	**chrF2++**	**BLEU**	**chrF2++**	**BLEU**	**chrF2++**	**BLEU**	**chrF2++**
S0 (Baseline)	41.2	60.3	**50.0**	**71.0**	**50.8**	**69.3**	41	61.1
S1	37.8	57.3	41.9	65.1	47.1	66.7	43.5	63.1
S2	35.7	56.9	46.1	67.4	45.6	66.7	42.3	63.1
S3	30.9	55.5	41.6	65.9	39.6	64	33.6	58.6
S4	25.3	52	39.8	64.7	38.3	63.1	28.8	55.8
S5	38.5	58.1	48.2	69.2	47.5	66.9	45.7	64.7
S6	40.2	59.4	49.1	70	49.2	68.3	45.4	64.4
S7 (Oracle)	**41.4**	**60.4**	49.5	70.9	50.5	69.1	45.4	64.4
S8 (Oracle-FT)	40.1	59.4	48.5	69.8	49	68.1	**46.1**	**65**

6. END-TO-END MACHINE TRANSLATION EVALUATION

The COVID-19 pandemic proved to be a crisis of extended duration and, therefore, apart from the domain adaptation experiments that we conducted in the previous chapter, we decided to further exploit the additional resources and experience that was gained in order to construct an end-to-end NMT system for EN→EL that could be embedded in a Crisis MT infrastructure (see Section 2.3.1). The implementation required significant changes compared to the defined experimentation framework in sections 5.1 and 5.2, as we aim to build a model which is robust to real-world cases and, due to hardware and time constraints, the system is only compared with available translation services and models. We perform limited experiments and focus on evaluation, since the process of creating a competitive end-to-end system is open-ended and requires more data, continuous changes, troubleshooting, and developments. Research and technical papers which focus on various components of NMT systems are published at a very fast rate and, consequently, an ablation study (or sensitivity analysis) on each component of such a system would require enormous hardware and time resources (tens or hundreds of GPUs); concerns about costly and environmentally unfriendly practices should also be raised and addressed by researchers.

6.1 Changes on the Training Configuration

Similarly with the configuration used in section (see Section 5.2), we utilize the Fairseq toolkit [90] to train big Transformer models [33] with a single RTX 2080 Ti GPU. However, we increase the number of encoder layers from 6 to 7 and the dimensionality of the inner layers of the feed-forward network from 4,096 to 5,120, as deeper and wider networks have been shown to improve performance in certain cases, especially with high data availability [92].

In the experiments of chapter 5, we trained a joint BPE tokenizer for both languages with 32,000 merge operations. This proved to be inefficient as it increased the parameters of the model, while further manual inspection revealed that it led to certain subwords being underrepresented in the training set. Therefore, in the experiments described here, we trained two independent tokenizers with 20,000 merge operations separately for each language. This resulted in an English dictionary with 20,252 subwords and a Greek dictionary with 20,236 subwords. Thus, the dimensionality of the embedding vectors of the encoder changed to 20,252 and to 20,236 for the decoder. Additionally, BPE-dropout with probability 0.1 was applied on the source (English) sentences, as it has been shown to improve robustness to typing errors and real-world applications [39, 94].

The rest of the architectural configurations are the same with section 5.2, except for the warm steps which were increased from 4,000 to 8,000. Nevertheless, there are some notable changes in the training set, its processing, and the methods that are used to improve the performance of the NMT systems.

The “inline casing” technique, which has been shown to be the optimal approach for handling casing [39, 40], was utilized through the application of three tags in the (lower-cased) training set: The <UC> tag was placed before uppercase subwords, the <TC> tag was used to denote title-cased subwords, and the <MC> tag denoted that the casing alternated in mixed-case subwords. In this way, these three tags are generated by the decoder during inference and are used to re-case the output sentences. This method also has the advantage of “mimicking” the way in which casing is applied on (user-given) inputs; for example, if an input sentence in English is constituted solely by capital letters, then the letters of the translated sentence in Greek will also be in the upper case.

We also utilize tagged back-translation (see Section 2.2.2) to improve performance on in-domain data [39, 45], by using another type of tag, <BT>, at the beginning of back-translated source sentences. In the implementation of the end-to-end NMT system, we gathered monolingual health-related corpora in Greek and translated them using the SPC-TUC EL→EN system [71] implemented in the Hugging Face library [70]. More specifically, we back-translated 27,550,834 Greek sentences (BackTranslatedLarge), as mentioned in section 3.2.4. After filtering and deduplication, 11,309,535 synthetic sentence pairs remained and were subsequently used in training, as listed in Table 6.

Similarly with the CovidWebsites corpus (see Section 3.2.3), we mined 508,332 additional sentence pairs from health-related websites, such as websites of hospitals, medical clinics, medical associations, and medical journals, and used them for training the end-to-end NMT system, alongside with the 11,309,535 synthetic sentence pairs and the 30,842,894 -originally- parallel sentences used in strategy S7 of the experiments in chapter 5. Thus, the training set of the E2E (end-to-end) system comprises 42,660,761 sentence pairs. We also fine-tune (E2E-FT) with 3,169,903 sentence pairs: the 2,661,571 that were used in strategy S6 (see Table 4) and the abovementioned 508,332 additional health-related parallel sentences that we mined. These sets are listed in Table 6.

Table 6: Description and size of the parallel data used in training the end-to-end NMT systems

	Type of Data	Number of sentence pairs used for training
(1)	Original Parallel Data (S7)	30,842,894
(2)	BackTranslatedLarge	11,309,535
(3)	Additional Health-Related Data	508,332
(4)	Fine-Tuning Data (S6)	2,661,571
E2E	(1) + (2) + (3)	42,660,761
E2E-FT	(3) + (4)	3,169,903

6.2 Results and Comparison

The E2E and E2E-FT systems that we have trained are compared with three publicly available systems and two commercial services. In particular, we compare with the EN→EL OPUS-MT system from Helsinki-NLP [95], its bigger variant (Helsinki OPUS-MT Big), SSE-TUC (Penelopie) [71], as well as the commercial services provided by Google Cloud (NMT v3 model) [68] and DeepL [96] through their respective APIs.

We believe that the results listed here give a good indication of the strengths and weaknesses of each translation system/service, since we use 8 distinct test sets of various domains with a total of 35,842 sentence pairs (see Table 3) for evaluation. Nevertheless, it is very important to note that our systems are the only ones which are guaranteed to not have been trained with any of the data on the test sets; even parallel data which contains a near duplicate sentence on any of its sides (see Section 3.3.1). We should also note that testing with the commercial services (Google Cloud, DeepL) was performed in mid-2022

and that it is quite possible that actual performance of those services may change or has already changed, as well as that the NMT v3 model provided through the Google Cloud Translation API is different from the Google Translate web service.

Table 7: Comparison of systems on COVID-19 generalization set and average performance on the other seven test sets

System	COVID-19 Generalization Set			Average Metrics for 7 other Test Sets		
	BLEU	**chrF2++**	**COMET**	**BLEU**	**chrF2++**	**COMET**
Helsinki OPUS-MT	36.9	58.2	70.78	38.69	60	77.72
Helsinki OPUS-MT Big	39	59.7	73.82	40.84	61.97	81.95
SSE – TUC (Penelopie)	29.1	51.7	47.34	34.30	56.56	64.37
Google NMT v3	39.3	60.9	79.64	40.96	62.33	86.65
DeepL	42.3	62.3	81.81	42.17	**63.44**	**88.91**
E2E	43.5	63	82.92	42.03	62.87	85.22
E2E-FT	**44.2**	**63.5**	**83.96**	**42.19**	62.81	85.86

In Table 7, we list the BLEU, chrF2++, and COMET metrics for the COVID-19 generalization set, as well as their -unweighted- average values for the other seven test sets that were used for evaluation. According to [80], it is better to give more importance to COMET scores, followed by chrF2++, and then BLEU. The following can be observed:

- Our systems, and in particular our fine-tuned system (E2E-FT), exhibit the best performance on the COVID-19 generalization set. The models by Google and DeepL are the most competitive, while the Helsinki OPUS-MT Big model seems to exhibit good translation quality as well. Despite the similar BLEU score of OPUS-MT Big (39) with that of Google Cloud NMT v3 (39.3), the differences in the chrF2++ (-1.2) and COMET (-5.82) metrics show that Google fairs better.
- Based on the average metrics for the seven other test sets, the DeepL model seems to exhibit the best general performance. Our fine-tuned system (E2E-FT) and Google NMT v3 have similar scores (+1.23 BLEU, +0.48 chrF2++, -0.79 COMET) and, as we will see later, E2E-FT shows better results on the test sets created in this thesis. Therefore, we should assume that our models (and Helsinki OPUS-MT Big to a lesser degree) are quite competitive but do not have as good general performance as the commercial translation services.
- The phenomenon of “catastrophic forgetting”, does not seem to be an issue at all, as the average metrics of the E2E and the E2E-FT models on the seven other test sets are very close. Thus, even though the improvements on the COVID-19 generalization set after fine-tuning seem minor (+0.7 BLEU, +0.5 chrF2++, +1.04 COMET), they do not incur an analogous degradation in performance for out-of-domain texts.

- The Helsinki OPUS-MT uses the Transformer-Align architecture [31], while the Helsinki OPUS-MT Big uses the big Transformer architecture from [33] and has been trained with additional back-translated data. As expected, these changes lead to better scores (especially regarding COMET) on all the eight test sets.

Table 8: Comparison of systems on seven other test sets

Test Set	Metrics	System						
		OPUS-MT	OPUS-MT Big	SSE-TUC	Google	DeepL	E2E	E2E-FT
General	BLEU	37.1	38.4	32.3	37.2	38.4	**40.6**	40.3
	chrF2++	57.5	58.8	53.9	58.3	59.2	**60.4**	59.9
	COMET	62.05	66.27	45.49	67.22	**70.13**	69.79	69.41
Tatoeba	BLEU	54.4	54.7	51.9	53.3	**55.6**	53.8	53
	chrF2++	72	72.4	70.3	71.4	**73.4**	72.2	71.6
	COMET	115.82	115.27	113	115.84	**117.17**	114.90	115.40
XNLI	BLEU	46.7	49.7	41.2	51.7	49.3	**52.2**	51.8
	chrF2++	65.9	68.3	61.7	69.9	68.6	**70.1**	69.6
	COMET	85.78	89.67	73.58	**98.42**	97.88	93.21	94.31
Ted Talks	BLEU	34.7	36.4	31.7	36.4	**37.6**	36.6	36.6
	chrF2++	57.4	59.7	55.2	59.8	**61.5**	59.8	59.2
	COMET	70.47	75.01	60.23	78.88	**84.58**	76.95	76.62
ACCURAT	BLEU	47.2	51.4	39.4	51.1	51.5	51.9	**52.1**
	chrF2++	67.9	70.9	62.1	71.1	**71.2**	71.1	**71.2**
	COMET	92.77	99.01	73.13	**103.45**	101.75	100.98	99.02
SciPar	BLEU	32.3	35.5	26.5	38	**39.7**	38.2	39
	chrF2++	56.9	59.8	52.2	62.2	**63.6**	62.2	62.8
	COMET	72.48	79.41	52.69	89.07	**90.19**	88.93	90.09
Literature	BLEU	18.4	19.8	17.1	19	**23.1**	20.9	22.5
	chrF2++	42.4	43.9	40.5	43.6	**46.6**	44.3	45.4
	COMET	44.65	49.02	32.48	53.70	**60.67**	51.76	56.17

In Table 8, we also list the metrics of each system for the seven other test sets that were used in the evaluation. The general remarks that can be drawn are:

- Although we mentioned before that the DeepL model exhibits the best general performance on these seven test sets, it is surpassed -in terms of COMET- by Google Cloud NMT v3 on the XNLI (+0.54 COMET) and ACCURAT (+1.7 COMET) test sets. Furthermore, in terms of BLEU and chrF2++ it is surpassed by our E2E system on the General (+2.2 BLEU, +1.2 chrF2++) and XNLI (+2.9 BLEU, +1.5 chrF2++) test sets (differences on ACCURAT are not significant). These results indicate that different metrics may lead to different conclusions when comparing MT systems, as well as that actual performance may be more similar than originally thought. For these reasons, it is prudent to evaluate and compare different systems as extensively as possible, in order to gain a better picture.
- All systems exhibit very high scores on the Tatoeba test set which is constituted mostly by generic and short sentences, while many of them differ slightly. The Helsinki OPUS-MT Big system outperforms -significantly- its simpler and older variant (Helsinki OPUS-MT) on all metrics and test sets, except the Tatoeba test set. Therefore, we believe that any automatic evaluation on the Tatoeba test set alone would not be very reliable and it is quite probable that this applies to other language pairs as well.
- The OPUS-MT Big system shows a significantly lower performance than our E2E-FT system on the SciPar (-3.5 BLEU, -3 chrF2++, -10.68 COMET) and Literature (-2.7 BLEU, -1.5 chrF2++, -7.15 COMET) test sets, i.e., domain-specific test set that have been created in this thesis. Moreover, it shows very similar performance on the Ted Talks and ACCURAT test sets; it does so on Tatoeba as well, although all systems do. Thus, the OPUS-MT Big system and our systems seem to exhibit quite comparable performance on general texts, although our systems seem to handle specific domains and genres better.
- The largest differences in terms of COMET of our systems from the best-performing ones seem to concern the XNLI, Ted Talks, and Literature test sets, i.e., test sets which contain dialogues, transcripts, spoken language, etc. We suspect that this may mean that our systems exhibit lower performance on these specific types of language, but further assessment is needed.

We should note that no comparison of MT systems can be conclusive, especially if automatic metric scores are similar or conflicting. Even a more time-consuming and detailed manual evolution would focus on specific use-cases or domains. Additionally, the comparison of a system with a commercial service, which is not entirely open to the public and may be updated without notice, is always less reliable. However, we believe that in-depth comparisons, such as the one that we conducted in this section, are very useful when building NMT systems, as they can reveal strengths and weaknesses which could be mitigated in a realistic manner. In connection with Crisis MT, an extensive evaluation framework could help us decide which pre-existing model is better suited to be adapted for the translation needs of a crisis scenario, as well as when to deploy our domain-adapted NMT model in a rapid response infrastructure (see Section 2.3.1).

7. CONCLUSION AND FUTURE WORK

In the present thesis, we investigate several components of Crisis MT, by focusing on the case of COVID-19 for English to Greek translations. We discuss the importance of MT in responding rapidly to crisis scenarios, as well as theoretical, practical, and philosophical issues related to Crisis MT and domain adaptation. The main contribution of this thesis is threefold:

- Describing the process of constructing high-quality and domain-specific parallel corpora from several sources. We use various tools and techniques to create corpora related to COVID-19 [7], scientific research [6], as well as additional originally parallel and synthetic corpora which may prove valuable in effective domain adaptation. We provide technical details on parallel data acquisition, filtering, pre-processing, augmentation, and selection.
- Experimenting with different rapid domain adaptation strategies in a simulated crisis scenario with variable data availability over time. These experiments focus on the fine-tuning order for adapting pre-existing NMT models (which we train from scratch) and the effect of "catastrophic forgetting", i.e., the decrease in performance on out-of-domain texts. Our results indicate that a single step of fine-tuning with in-domain and close-to-domain data leads to performance similar with a domain-adapted oracle model, as well as that mixed fine-tuning, even on a limited scale, can improve the balance between performance on out-of-domain and in-domain texts.
- Building an end-to-end NMT system (using the Transformer architecture) for the English→Greek translation direction with the use of data augmentation and other techniques which improve robustness to real-world applications. We compare its performance with other publicly available models and translation services on various test sets, some of which are novel and were constructed as part of this thesis. Our domain-adapted system seems to achieve the highest performance on COVID-19 related texts, while also exhibiting a very competitive performance on other domains.

These contributions hopefully provide useful insights on practical aspects of Crisis MT, such as data acquisition and selection, the choice of the domain adaptation strategy (e.g., fine-tuning order, mixed fine-tuning), and the evaluation framework which can help us decide which pre-existing system should be used for fine-tuning and when should the domain-adapted system be deployed in a rapid response infrastructure.

In the future, we aim to also construct a NMT system for the Greek→English translation direction and language models for Greek and English, which will be used alongside the EN→EL system for the construction of state-of-the-art NMT systems in the biomedical and scientific domains, utilizing additional techniques which have been shown to lead to better translation quality (e.g., iterative back-translation, reranking, domain control). Furthermore, we plan to experiment with more domain adaptation strategies (e.g., mixed fine-tuning, back-translation) and assess their efficacy in other lower- and higher-resource MT scenarios and other domains, such as other crises scenarios.

In the long term, we plan on further investigating the key components of Crisis MT towards proposing a more complete course of action in crisis scenarios which will also involve effective data acquisition, selection, augmentation, and filtering. Hopefully, we will be able to map the steps on delivering expert solutions for researchers, translation professionals, private entities, and the public, with a focus on scenarios in which a specific need has arisen and must be addressed promptly and adequately.

ABBREVIATIONS - ACRONYMS

AI	Artificial Intelligence
API	Application Programming Interface
BPE	Byte Pair Encoding
CORD-19	Covid-19 Open Research Dataset
COVID-19	Coronavirus Disease 2019
CTD	Close-to-domain
ECDC	European Centre for Disease Prevention and Control
EU	European Union
ILSP-FC	Institute of Language and Speech Processing – Focused Crawler
IND	In-domain
LASER	Language Agnostic Sentence Representations
MT	Machine Translation
NLP	Natural Language Processing
NMT	Neural Machine Translation
OOD	Out-of-domain
PDF	Portable Document Format
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
TMX	Translation Memory Exchange
URL	Uniform Resource Locator
UTF	Unicode Transformation Format
WHO	World Health Organization
XML	Extensible Markup Language

APPENDIX

A.1 List of terms used to determine strict “covidity”

'covid', 'covid-19', 'pandemic', 'coronavirus', 'koronavirus', 'sars-cov-2'

A.2 List of terms used to determine extended “covidity”

'adenovirus', antibod', 'antigen', 'antimicrob', 'antiseptic', 'antiviral', 'astrazeneca', 'asymptomatic', 'biological', 'biopolitic', 'blood clot', 'border test', 'care worker', 'cases confirmed', 'chloroquine', 'clinic', 'community spread', 'confirmed case', 'confirmed death', 'contact tracing', 'contagio', 'contaminat', 'coronavac', 'covax', 'crisis', 'death toll', 'deaths confirmed', 'dexamethasone', 'diagnosis', 'diagnostic test', 'disease', 'disinfectant', 'distancing', 'doctor', 'economical impact', 'epidemic', 'epidemiological', 'essential worker', 'fatality ratio', 'fever', 'gathering restriction', 'genome', 'green pass', 'health', 'herd immunity', 'hospital', 'hydroxychloroquine', 'icu beds', 'illness', 'immune system', 'immunity', 'immunization', 'immunosuppression', 'index case', 'infect', 'inflammatory', 'influenza', 'intensive care unit', 'johnson', 'key worker', 'lockdown', 'mandatory measure', 'mandatory prohibition', 'mandatory restriction', 'martial law', 'masks', 'medical', 'medication', 'medicine', 'mers cov', 'microb', 'moderna', 'mordibidity', 'mortality rate', 'mrna', 'mutation', 'myocarditis', 'new cases', 'new recoveries', 'new strain', 'olfactory', 'operation freedom', 'outbreak', 'pasteur', 'pathogen', 'patient', 'patient zero', 'pcr test', 'pfizer', 'physician', 'pneumonia', 'prohibitive measure', 'quarantine', 'rapid test', 'reinfection', 'remdesivir', 'remote work', 'repatriation', 'reproduction number', 'reproduction rate', 'respiratory', 'restrictive measure', 'sanita', 'sanitizer', 'sars cov', 'self isolation', 'self test', 'side effect', 'sinopharm', 'sinovac', 'social distanc', 'social impact', 'spike protein', 'spreader', 'sputnik', 'stay home', 'superspreader', 'symptom', 'syncytial', 'syndrome', 'teleworking', 'temporary measure', 'thrombosis', 'transmissibility', 'transmission', 'travel restriction', 'trial subject', 'vaccin', 'variant', 'ventilated', 'ventilator', 'viral vector', 'virus', 'wash hands', 'who directives', 'who guidelines'

REFERENCES

[1] J. Shuja, E. Alanazi, W. Alasmary and A. Alashaikh, "COVID-19 open source data sets: a comprehensive survey," *Applied Intelligence,* vol. 51, pp. 1296-1325, 2021.

[2] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney and others, "CORD-19: The COVID-19 Open Research Dataset," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.

[3] D. Saunders, "Domain adaptation and multi-domain adaptation for neural machine translation: A survey," *Journal of Artificial Intelligence Research,* vol. 75, pp. 351-424, 2022.

[4] W. Lewis, R. Munro and S. Vogel, "Crisis MT: Developing a cookbook for MT in crisis situations," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011.

[5] D. Roussis and V. Papavassiliou, "The ARC-NKUA submission for the English-Ukrainian General Machine Translation Shared Task at WMT22," in *Proceedings of the Seventh Conference on Machine Translation*, 2022.

[6] D. Roussis, V. Papavassiliou, P. Prokopidis, S. Piperidis and V. Katsouros, "SciPar: A Collection of Parallel Corpora from Scientific Abstracts," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022.

[7] D. Roussis, V. Papavassiliou, S. Sofianopoulos, P. Prokopidis and S. Piperidis, "Constructing Parallel Corpora from COVID-19 News using MediSys Metadata," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022.

[8] M. Verile, G. Jacquet, L. Della Rocca and E. Mantica, "COVID-19 news monitoring with Medical Information System (Medisys)," 2020.

[9] F. Casacuberta, A. Ceausu, K. Choukri, M. Deligiannis, M. Domingo, M. Garcıa-Martınez, M. Herranz, V. Papavassiliou, S. Piperidis, P. Prokopidis and others, *The Covid-19 MLIA@ Eval initiative: Overview of the machine translation task,* 2021.

[10] F. Casacuberta, A. Ceausu, K. Choukri, M. Deligiannis, M. Domingo, M. García-Martínez, M. Herranz, G. Jacquet, V. Papavassiliou, S. Piperidis, P. Prokopidis, D. Roussis and M. Hadj Salah, "Findings of the Covid-19 MLIA Machine Translation Task," *arXiv preprint arXiv:2211.07465,* 2022.

[11] A. Anastasopoulos, A. Cattelan, Z.-Y. Dou, M. Federico, C. Federmann, D. Genzel, F. Guzmán, J. Hu, M. Hughes, P. Koehn and others, "TICO-19: the Translation Initiative for COvid-19," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

[12] ParaCrawl, "ParaCrawl Synthesized Data," https://paracrawl.eu/manufactured-data.

[13] SYSTRAN, "12 Translation models specialized with Corona Crisis Data," https://www.systransoft.com/systran/news-and-events/specialized-corona-crisis-corpus-models/.

[14] L. Yeganova, D. Wiemann, M. Neves, F. Vezzani, A. Siu, I. J. Unanue, M. Oronoz, N. Mah, A. Neveol, D. Martinez and others, "Findings of the WMT 2021 Biomedical Translation Shared Task: Summaries of Animal Experiments as New Test Set," in *Proceedings of the Sixth Conference on Machine Translation*, 2021.

[15] M. M. ibn Alam, I. Kvapilíková, A. Anastasopoulos, L. Besacier, G. Dinu, M. Federico, M. Gallé, P. Koehn, V. Nikoulina and K. W. Jung, "Findings of the WMT Shared Task on Machine Translation Using Terminologies," in *Proceedings of the Sixth Conference on Machine Translation*, 2021.

[16] A. K. Ojha, C.-H. Liu, K. Kann, J. Ortega, S. Shatam and T. Fransen, "Findings of the LoResMT 2021 Shared Task on COVID and Sign Language for Low-resource Languages," in *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, 2021.

[17] M. Mahdieh, M. X. Chen, Y. Cao and O. Firat, "Rapid domain adaptation for machine translation with monolingual data," *arXiv preprint arXiv:2010.12652,* 2020.

[18] S. Bandyopadhyay, "Factored Neural Machine Translation on Low Resource Languages in the COVID-19 crisis," 2020.

[19] A. Way, R. Haque, G. Xie, F. Gaspari, M. Popović and A. Poncelas, "Rapid development of competitive translation engines for access to multilingual COVID-19 information," in *Informatics*, 2020.

[20] Z. Hou, F. Du, X. Zhou, H. Jiang, S. Martin, H. Larson and L. Lin, "Cross-Country Comparison of Public Awareness, Rumors, and Behavioral Responses to the COVID-19 Epidemic: Infodemiology Study," *Journal of Medical Internet Research,* vol. 22, 2020.

[21] C. Cuello-Garcia, G. Pérez-Gaxiola and L. van Amelsvoort, "Social media can have an impact on how we manage and investigate the COVID-19 pandemic," *Journal of clinical epidemiology,* vol. 127, pp. 198-201, 2020.

[22] P. E. Skarpa and E. Garoufallou, "Information seeking behavior and COVID-19 pandemic: A snapshot of young, middle aged and senior individuals in Greece," *International Journal of Medical Informatics,* vol. 150, p. 104465, 2021.

[23] L. Papadopoulou and T. A. Maniou, "'Lockdown'on Digital Journalism? Mapping Threats to Press Freedom during the COVID-19 Pandemic Crisis," *Digital Journalism,* vol. 9, pp. 1344-1366, 2021.

[24] W. V. O. Quine, Word and object, 2013.

[25] W. V. Quine, "On the reasons for indeterminacy of translation," *The Journal of Philosophy,* vol. 67, pp. 178-183, 1970.

[26] L. Wittgenstein, Philosophical investigations, John Wiley & Sons, 2010.

[27] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781,* 2013.

[28] T. Mikolov, W.-t. Yih and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 2013.

[29] K. Cho, B. van Merriënboer, D. Bahdanau and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014.

[30] D. Saunders, "Domain adaptation for neural machine translation," 2021.

[31] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.

[32] M.-T. Luong, H. Pham and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[34] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[35] J. L. Ba, J. R. Kiros and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450,* 2016.

[36] R. Sennrich, B. Haddow and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[37] P. Gage, "A new algorithm for data compression," *C Users Journal,* vol. 12, pp. 23-38, 1994.

[38] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the ACL 2007 Demo and Poster Sessions*, 2007.

[39] A. Bérard, I. Calapodescu and C. Roux, "Naver Labs Europe's Systems for the WMT19 Machine Translation Robustness Task," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 2019.

[40] T. Etchegoyhen and H. Gete, "To case or not to case: Evaluating casing methods for neural machine translation," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020.

[41] A. Molchanov, "PROMT Systems for WMT 2020 Shared News Translation Task," in *Proceedings of the Fifth Conference on Machine Translation*, 2020.

[42] C. Chu, R. Dabre and S. Kurohashi, "An empirical comparison of domain adaptation methods for neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017.

[43] R. Sennrich, B. Haddow and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[44] S. Edunov, M. Ott, M. Auli and D. Grangier, "Understanding Back-Translation at Scale," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[45] I. Caswell, C. Chelba and D. Grangier, "Tagged Back-Translation," in *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 2019.

[46] Qatar Computing Research Institute and Hamad Bin Khalifa University, "Resources for Research on Crisis Informatics Topics," https://crisisnlp.qcri.org/.

[47] J. Tiedemann, "Parallel Data, Tools and Interfaces in OPUS," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

[48] V. Papavassiliou, P. Prokopidis and G. Thurmair, "A modular open-source focused crawler for mining monolingual and bilingual corpora from the web," in *Proceedings of the sixth workshop on building and using comparable corpora*, 2013.

[49] M. Artetxe and H. Schwenk, "Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[50] R. Steinberger, A. Eisele, S. Klocek, S. Pilos and P. Schlüter, "DGT-TM: A freely available Translation Memory in 22 languages," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

[51] R. Skadiņš, J. Tiedemann, R. Rozis and D. Deksne, "Billions of parallel words for free: Building and using the eu bookshop corpus," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

[52] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *The Tenth Machine Translation Summit Proceedings of Conference*, 2005.

[53] F. M. Tyers and M. S. Alperen, "South-east european times: A parallel corpus of balkan languages," in *Proceedings of the LREC workshop on exploitation of multilingual resources and tools for Central and (South-) Eastern European Languages*, 2010.

[54] K. Wołk and K. Marasek, "Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs," *Procedia Technology,* vol. 18, pp. 126-132, 2014.

[55] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.

[56] P. Lison, J. Tiedemann and M. Kouylekov, "Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[57] N. Reimers and I. Gurevych, "Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[58] H. Schwenk, V. Chaudhary, S. Sun, H. Gong and F. Guzmán, "WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021.

[59] A. El-Kishky, V. Chaudhary, F. Guzmán and P. Koehn, "CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[60] H. Schwenk, G. Wenzek, S. Edunov, É. Grave, A. Joulin and A. Fan, "CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

[61] V. Papavassiliou, P. Prokopidis and S. Piperidis, "Discovering parallel language resources for training MT engines," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[62] European Centre for Disease Prevention and Control, "External resources on COVID-19," https://www.ecdc.europa.eu/en/covid-19/external-resources.

[63] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary and others, "Beyond english-centric multilingual machine translation," *Journal of Machine Learning Research,* vol. 22, pp. 1-48, 2021.

[64] B. Thompson, "Vecalign," https://github.com/thompsonb/vecalign.

[65] European Union, "European Medicines Agency," https://www.ema.europa.eu/.

[66] European Language Resources Coordination, "ELRC-SHARE Repository," https://elrc-share.eu/.

[67] Covid-19 MLIA - Eval, "Covid-19 MLIA - Eval Tasks," http://eval.covid19-mlia.eu/.

[68] Google, "Google Cloud Translation AI," https://cloud.google.com/translate.

[69] CLARIN:EL, "Central Inventory of Language Resources and Services," https://inventory.clarin.gr/.

[70] Hugging Face, "Hugging Face Models," https://huggingface.co/models.

[71] D. Papadopoulos, N. Papadakis and N. Matsatsinis, "PENELOPIE: Enabling Open Information Extraction for the Greek Language through Machine Translation," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 2021.

[72] M. Pinnis, "Tilde's parallel corpus filtering methods for WMT 2018," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 2018.

[73] V. Papavassiliou, S. Sofianopoulos, P. Prokopidis and S. Piperidis, "The ILSP/ARC submission to the WMT 2018 Parallel Corpus Filtering Shared Task," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 2018.

[74] M. Rikters, "Impact of Corpora Quality on Neural Machine Translation," in *Human Language Technologies–The Baltic Perspective*, IOS Press, 2018, pp. 126-133.

[75] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, "Bag of Tricks for Efficient Text Classification," *arXiv preprint arXiv:1607.01759,* 2016.

[76] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002.

[77] B. Marie, A. Fujita and R. Rubino, "Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

[78] M. Post, "A Call for Clarity in Reporting BLEU Scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018.

[79] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 2015.

[80] T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita and A. Menezes, "To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation," in *Proceedings of the Sixth Conference on Machine Translation*, 2021.

[81] M. Popović, "chrF deconstructed: beta parameters and n-gram weights," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 2016.

[82] M. Popović, "chrF++: words helping character n-grams," in *Proceedings of the Second Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*, 2017.

[83] R. Rei, C. Stewart, A. C. Farinha and A. Lavie, "COMET: A Neural Framework for MT Evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[84] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *Advances in neural information processing systems,* vol. 32, 2019.

[85] J. Tiedemann, "The Tatoeba Translation Challenge–Realistic Data Sets for Low Resource and Multilingual MT," in *Proceedings of the Fifth Conference on Machine Translation*, 2020.

[86] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk and V. Stoyanov, "XNLI: Evaluating Cross-lingual Sentence Representations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[87] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan and G. Neubig, "When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation?," in *Proceedings of the 2018 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018.

[88] CLARIN:EL, "ACCURAT Balanced Test Corpus for Under Resourced Languages," https://inventory.clarin.gr/corpus/662, 2010.

[89] W. Wang, W. Peng, X. Meng and Q. Liu, "Huawei AARC's Submissions to the WMT21 Biomedical Translation Task: Domain Adaption from a Practical Perspective," in *Proceedings of the Sixth Conference on Machine Translation*, 2021.

[90] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier and M. Auli, "FAIRSEQ: A Fast, Extensible Toolkit for Sequence Modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[91] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* 2014.

[92] S. Subramanian, O. Hrinchuk, V. Adams and O. Kuchaiev, "NVIDIA NeMo's Neural Machine Translation Systems for English-German and English-Russian News and Biomedical Tasks at WMT21," in *Proceedings of the Sixth Conference on Machine Translation*, 2021.

[93] C. Kobus, J. M. Crego and J. Senellart, "Domain Control for Neural Machine Translation," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 2017.

[94] I. Provilkov, D. Emelianenko and E. Voita, "BPE-Dropout: Simple and Effective Subword Regularization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[95] J. Tiedemann and S. Thottingal, "OPUS-MT–Building open translation services for the World," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 2020.

[96] DeepL, "DeepL Translator API," https://www.deepl.com/pro-api?cta=header-pro-api.